Πανεπιστημιο Θεσσαλιας

Μεταπτυχιακη εργασια

Ακουστική Ανάλυση Σκηνής σε Πολυκαναλικά Περιβάλλοντα

Συγγραφέας:
Κωνσταντίνος Α. Θεμελής

Επιβλέποντες:
Γεράσιμος Ποταμιάνος
Αντώνιος Αργυρίου
Νικόλαος Μπέλλας

Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

3 Ιουλίου 2018

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 27 Ιουνίου 2018.

Γεράσιμος Ποταμιάνος,
Αναπληρωτής Καθηγητής
Υπογραφή:

Αργυρίου Αντώνιος,
Επίκουρος Καθηγητής
Υπογραφή:

Μπέλλας Νικόλαος,
Αναπληρωτής Καθηγητής
Υπογραφή:

Πανεπιστήμιο Θεσσαλίας

Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Περίληψη

Η ακουστική ανάλυση σκηνής στοχεύει στην επεξεργασία και την ερμηνεία της ηχητικής πληροφορίας που διαδόθηκε στο περιβάλλον και ηχογραφήθηκε από πολύ-μικροφωνικές συστοιχίες. Η αναγνώριση και η ταξινόμηση των ακουστικών γεγονότων με χρήση αναπτυγμένων αλγορίθμων και μοντέλων αποτελούν βασικά ερευνητικά ζητήματα για την ακουστική σκηνή. Υπάρχουν πολλοί ταξινομητές όπως το πολυεπίπεδο perceptron (MLP multilayer perceptron-MLP), HMM (hidden Markov models), Bayes κτλπ, αλλά στο πλαίσιο αυτής της εργασίας θα πειραματιστούμε με έναν συνδυασμό νευρωνικών δικτύων και κρυφών μοντέλων Markov, δημιουργώντας έτσι ένα υβριδικό νευρωνικό δίκτυο.

Κατά τη διάρκεια του σχεδίου εργασίας, χρησιμοποιήσαμε την πολυκαναλική βάση δεδομένων του ερευνητικού κέντρου UPC-TALP που ηχογραφήθηκε από 24 μικρόφωνα τοποθετημένα σε ένα έξυπνο περιβάλλον. Τα αρχεία έχουν τμηματοποιηθεί ήδη από προηγούμενη εργασία. Το βασικό εργαλείο που χρησιμοποιήσαμε είναι το HTK για να μοντελοποιήσουμε, να εκπαιδεύσουμε και να εξετάσουμε τη βάση δεδομένων με χρήση βαθιών νευρωνικών δικτύων. Στο τέλος, συγκρίναμε τα αποτελέσματα και διακρίναμε ποιά μέθοδος ανταποκρίνεται καλύτερα στο πρόβλημά μας.

University of Thessaly

Department of Electrical & Computer Engineering

Abstract

The acoustic scene analysis aims at processing and interpreting the audio information that was propagated in a environment and recorded by multiple microphones. Developing algorithms and models that recognize and classify a set of acoustic events is a research topic of high interest in the acoustic scene analysis. There are many classifiers such as perceptron-MLP, HMM (hidden Markov models), Bayes etc., but in the aspect of this thesis, we will experiment with a combination of neural networks and HMMs, creating a hybrid architecture.

During the work plan, we used the UPC-TALP multi-channel database of 24 microphones embedded in an intelligent environment. The files have already been documented by previous work. The basic tool we used for modeling, training and testing our system with neural networks is HTK. In the end, we compared the results of HMM and DNN to find which solution method suits better for our problem.

Ευχαριστίες

Κατ΄ αρχήν θα ήθελα να ευχαριστήσω τον επικεφαλή επόπτη αυτής της μεταπτυχιακής εργασίας κ.Γεράσιμο Ποταμιάνο, αναπληρωτή καθηγητή του τμήματος Ηλετρολόγων Μηχανικών & Μηχανικών Υπολογιστών στον Βόλο, για την παραπάνω από χρήσιμη βοήθειά του. Η καθοδήγησή του ήταν εξέχουσας σημασίας καθώς κατάφερα κατανοήσω βαθύτερα την επιστήμη των νευρωνικών δικτύων. Επιπρόσθετα, θα ήθελα να ευχαριστήσω από καρδιάς την οικογένειά και τους φίλους μου για την αγάπη και την υποστήριξή τους που με βοήθησε πνευματικά. Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στους αναπληρωτές καθηγητές Νικόλαο Μπέλλα και Αντώνιο Αργυρίου για την καθοδήγησή τους. Χωρίς τα προαναφερθέντα άτομα αυτή η μεταπτυχιακή εργασία δεν θα υπήρχε.

Στην οικογένειά μου και στους φίλους μου...

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

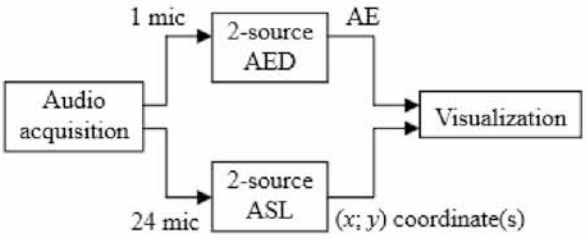
Εισαγωγή

1.1 Ακουστική ανάλυση σκηνής

Η ακουστικής ανάλυση σκηνής αναφέρεται γενικά στη διαδικασία στην οποία ένα ακουστικό σενάριο ανιχνεύεται ή ταξινομείται. Σε γενικές γραμμές, το ακουστικό σενάριο αποτελείται από ακουστικά γεγονότα που παράγονται από διαφορετικές πηγές, μέσα σε ένα περιβάλλον το οποίο είναι συνήθως δωμάτιο ή οποιοδήποτε άλλο αντήχιο περίβλημα. Η ανάλυση της ακουστικής σκηνής μπορεί να περιλαμβάνει την εκτίμηση ή την ταξινόμηση των παραμέτρων του ακουστικού περιβάλλοντος που συμβαίνει στο ακουστικό σενάριο. Το σήμα εισόδου σε ένα σχήμα ανάλυσης σκηνής μπορεί να είναι οποιαδήποτε μορφή σήματος που εκπέμπεται από ένα πηγή στο δωμάτιο, η οποία καταγράφηκε είτε από ένα μόνο μικρόφωνο, είτε από μια συστοιχία μικροφώνων[6]. Όπως γίνεται αντιληπτό, αυτός ο κλάδος μπορεί να λειτουργήσει σαν βελτιωτικός παράγοντας σε ταξινομητές που χρησιμοποιούνται για αναγνώριση φωνής [5]. Παρόλο που η ομιλία αποτελεί το πιο πληροφοριακό ακουστικό γεγονός, τα διαφορετικά γεγονότα που λαμβάνουν μέρος σε ένα χώρο μπορεί να φέρουν εξίσου σημαντική πληροφορία. Και αυτό γιατί μέσα από αυτά τα συμβάντα αντανακλάται η ανθρώπινη δραστηριότητα είτε από άμμεσες παρεμβάσεις στο γύρω περιβάλλον (ήχος βημάτων), είτε μέσα από τη χρήση αντικειμένων (ήχος πληκτρολογίου). Συνεπώς, με την αναγνώριση και κατηγοριοποίηση τέτοιων γεγονότων, μπορεί να χαρτογραφηθεί η ανθρώπινη και κοινωνική δραστηριότητα σε έξυπνα περιβάλλοντα [7]. Η αναγνώριση ακουστικών γεγονότων σε πολυκαναλικά περιβάλλοντα αποτελεί παρακλάδι της επιστήμης της Αναγνώρισης Προτύπων, καθώς τα ηχητικά σήματα επεξεργάζονται ώστε να εξαχθούν χαρακτηριστικά που μπορούν να τα περιγράψουν. Σύμφωνα με αυτά τα χαρακτηριστικά, εκπαιδεύονται ταξινομητές ώστε να καταστούν ικανοί να εντοπίσουν και να αναγνωρίσουν μεμονωμένα ακουστικά γεγονότα (μεμονωμένα πειράματα) ή ακολουθίες (ενσωματωμένα πειράματα).

1.2 Σχετική εργασία

Η μεταπτυχιακή αυτή εργασία βασίζεται στη δουλεία που ολοκληρώθηκε κατά τη δημιουργία της διπλωματικής εργασίας "Αναγνώριση ακουστικών γεγονότων με βαθιά νευρωνικά δίκτυα" [8]. Τόσο η θεωρία, όσο και το περιεχόμενο των ακουστικών δεδομένων παραμένουν στο ίδιο επίπεδο. Επιπρόσθετα, η βάση δεδομένων όπου πραγματοποιήθηκαν τα πειράματα, έχει χρησιμοποιηθεί σαν υλικό εκπαίδευσης και αξιολόγησης στο σχέδιο εργασίας "Multi-microphone fusion for detection of speech and acoustic events in smart spaces" [9]. Πειράματα πάνω στην ίδια ακριβώς βάση πραγματοποίησαν οι Taras Butko, Fran González Pla κ.α που προσπάθησαν να αναγνωρίσουν τα ακουστικά γεγονότα μέσω HMM ταξινομητή και να εντοπίσουν την ηχητική πηγή. Η ροή εργασίας που ακολούθησαν, δημιουργώντας ένα τετοιο σύστημα φαίνεται στο Σχήμα 1.1 [1].



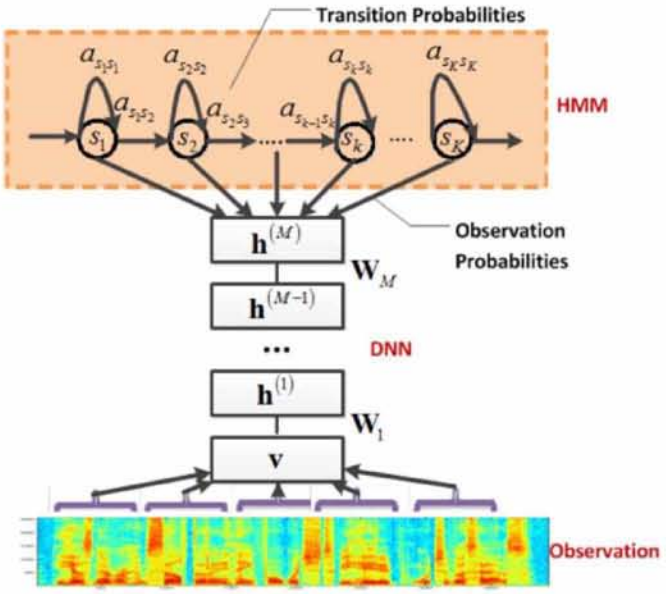
Σχημα 1.1: Αναπαράσταση του προτεινόμενου συστήματος. Λήφθηκε από το [1]

Πειράματα πάνω στα υβριδικά νευρωνικά δίκτυα πραγματοποίησαν οι George E. Dahl, Dong Yu [2] κ.α, αφού χρησιμοποίησαν ένα προ-εκπαιδευμένο contect-dependent νευρωνικό δίκτυο για να προσεγγίσουν το πρόβλημα της αναγνώρισης φωνής με μεγάλο λεξιλόγιο. Συγκεκριμένα, κατάφεραν με χρήση GPU να δημιουργήσουν ένα μοντέλο 5 κρυφών επιπέδων με 2000 κόμβους το καθένα και να επιτύχουν ακρίβεια στην αναγνώριση προτάσεων ύψους 70.3%.

1.3 Βαθιά μάθηση και νευρωνικά δίκτυα

Τα νευρωνικά δίκτυα είναι εμπνευσμένα από το βιολογικό μοντέλο που προτάθηκε από τον H. Hubel και Torsten Wiesel το 1959 [10]. Από βιολογική άποψη ο εγκέφαλος αποτελείται από νευρικά κύτταρα που ονομάζονται νευρώνες και είναι συνδεδεμένοι μεταξύ τους με δεσμούς. Αυτός ο μηχανισμός επιτρέπει τον ηλεκτρισμό να κυλλά διαμέσου του εγκεφάλου σε όλο το σώμα και έτσι καθιστά ικανή την επικοινωνία μεταξύ των κυττάρων. Έχοντας ο εγκέφαλος αυτόν τον μηχανισμό, λαμβάνει αποφάσεις και επιτρέπει σωματικές δράσεις. Τα τεχνητά νευρωνικά δίκτυα προσπαθούν να προσομοιώσουν αυτή τη συμπεριφορά δημιουργώντας ένα σύστημα κόμβων διαιρεμένο σε συνδεδεμένα επίπεδα. Στη συγκεκριμένη

εργασία, χρησιμοποποιήθηκε ένα είδος υβριδικού νευρωνικού δικτύου που περνάει ένα στάδιο προ-εκπαίδευσης με σκοπό την καλύτερη αρχικοποίηση των βαρών στα κρυφά επίπεδα. Στο Σχήμα 1.1 παρουσιάζεται παράδειγμα ενός νευρωνικού δικτύου όπου φαίνεται η οργάνωση των επιπέδων και οι συνδέσεις μεταξύ τους.



Σχημα 1.2: Διάγραμμα υβριδικής αρχιτεκτονικής. Το μέγεθος της εισόδου στο νευρωνικό δίκτυο συμπίπτει με τη διάσταση των χαρακτηριστικών.[2]

1.4 Η μεταπτυχιακή εργασία σε μια ματιά

1.4.1 Σκοπός της μεταπτυχιακής εργασίας

Η παρούσα εργασία αποτελεί επέκταση της μελέτη του προβλήματος της αναγνώρισης ακουστικών γεγονότων που λαμβάνουν χώρο σε ένα έξυπνο περιβάλλον. Τα γεγονότα έχουν ηχογραφηθεί ταυτόχρονα από 24 μικρόφωνα. Ο σκοπός αυτού του σχεδίου εργασίας είναι να μοντελοποιήσει το πρόβλημα με ένα υβριδικό νευρωνικό δίκτύο, να μελετήσει την επίδοσή τους μέσα από μία σειρά πειραμάτων, να εξάγει τα αποτελέσματα και τέλος να τα με αυτά των μαρκοβιανών μοντέλων. Επιπλέον, μέσω των πειραμάτων βλέπουμε την απόδοση του λογισμικού HTK σε σχέση με το Kaldi[8]. Πρέπει να σημειωθεί ότι τα δύο εργαλεία επιτρέπουν την μοντελοποίηση νευρωνικών δικτύων και τα αποτελέσματά τους θα παρουσιαστούν αναλυτικά στο Κεφάλαιο 5.

1.4.2 Συνεισφορά της εργασίας

Η κύρια συνεισφορά αυτής της μεταπτυχιακής εργασίας έγκειται στην εφαρμογή της μεθόδου των βαθιών υβριδικών νευρωνικών δικτύων στα δεδομένα της πολυκαναλικής βάσης δεδομένων UPC-TALP με σκοπό να εκτιμηθεί η αποδοτικότητα στην ανίχνευση των ακουστικών γεγονότων σε έξυπνα περιβάλλοντα. Πιο ειδικά, η επιστημονική συνεισφορά της εργασίας μπορεί να συνοψιστεί στους ακόλουθους τομείς:

- Τη χρήση διαφορετικών μετρικών λάθους με σκοπό την καλύτερη και ορθότερη αξιολόγηση των ταξινομητών.
- Εύρεση του αποδοτικότερου μοντέλου υβριδικού βαθιού νευρωνικού δικτύου για την μοντελοποίηση του προβλήματος.
- Εξαγωγή τελικών συμπερασμάτων και αξιολόγηση της εφαρμογής νευρωνικών δικτύων στην ηχογραφημμένη βάση δεδομένων.
- Αξιολόγηση του HTK ως εναλλακτικό λογισμικό δημιουργίας νευρωνικών δικτύων.
- Αξιολόγηση των χαρακτηριστικών MFCC ως προς την αποδοτικότητα της εφαρμογής του στα βαθιά νευρωνικά δίκτυα.
- Αξιολόγηση της επίδρασης των μοντέλων Markov στη δημιουργία ενός αποδοτικού υβριδικού νευρωνικού δικτύου.
- Σύγκριση μετρικών για την ανάδειξη του καλύτερου πιθανοτικού μοντέλου με βάση τα πειράματα.

1.4.3 Περιεχόμενο κεφαλαίων

Εκτός της εισαγωγής, αυτή η εργασία αποτελείται από πέντε ακόμα κεφάλαια που συμπεριλαμβάνουν το ακόλουθο περιεχόμενο:

- ***Κεφάλαιο 2***: περιέχει όλη την πληροφορία σχετικά με τον τρόπο επίλυσης του προβλήματος, μία συνοπτική αναφορά στα μοντέλα Markov, τη θεωρία όπου βασίζεται ο συνδυασμός τους με βαθιές αρχιτεκτονικές, την εξαγωγή των χαρακτηριστικών MFCC που χρησιμοποιήθηκαν για την αναγνώριση ακουστικών γεγονότων και τις ιδιότητές τους. Επιπλέον, παραθέτονται μαθηματικοί τύποι με σκοπό την πλήρη κατανόηση των παραπάνω διαδικασιών.
- ***Κεφάλαιο 3***: σε αυτό το τμήμα παρουσιάζεται η πολυκαναλική βάση δεδομένων UPC-TALP, τεχνικές πληροφορίες σχετικά με το υπόβαθρο στο οποίο ηχογραφήθηκε

και οργανώθηκε. Ακόμη, αναρτείται και πίνακας όπου τα δεδομένα της βάσης έχουν κατηγοριοποιηθεί στις 8 συνεδρίες ανάλογα με το γεγονός που περιγράφουν. Σε αυτό το σημείο πρέπει να σημειωθεί ότι το οπτικό τμήμα της βάσης δεν περιγράφεται καθώς δεν χρησιμοποιήθηκε στο πλαίσιο της εργασίας.

- ***Κεφάλαιο 4***: το πιο εκτενές και αναλυτικό κεφάλαιο της εργασίας, αφού περιγράφεται το βασικό εργαλείο-λογισμικό HTK που χρησιμοποιήθηκε για την διεκπεραίωση της εργασίας. Έπειτα, παραθέτονται αναλυτικές πληροφορίες για το εργαλείο SoX με το οποίο έγινε η επεξεργασία των ακουστικών αρχείων σε μία κοινή μορφή. Στο τέλος του κεφαλαίου, περιγράφεται η διαδικασία της εξαγωγής χαρακτηριστικών, η χρήση του καλύτερου HMM-GMM μοντέλου για την προεκπαίδευση και εκπαίδευση ενός νευρωνικού δικτύου. Τέλος, περιγράφεται η διαδικασία της αποκωδικοποίησης των δεδομένων με χρήση του υβριδικού μοντέλου DNN-HMM.

- ***Κεφάλαιο 5***: εδώ παρουσιάζονται σε πίνακες τα αποτελέσματα που εξήχθησαν από το HTK για την αναγνώριση των ακουστικών γεγονότων με το υβριδικό μοντέλο που δημιουργήσαμε. Υπάρχει πλήρης σχολιασμός και σύγκριση των αποτελεσμάτων καθώς και γραφήματα για την ευκολότερη κατανόησή τους.

- ***Κεφάλαιο 6***: στο τελευταίο κεφάλαιο πραγματοποιείται συζήτηση σχετικά με τα αποτελέσματα του Κεφαλαίου 5 και προτείνονται μελλοντικοί οδοί βελτίωσης των αποτελεσμάτων μέσα από την εξαγωγή διαφορετικών χαρακτηριστικών ή συνδυασμό καναλιών.

Κεφάλαιο 2

Χαρακτηριστικά MFCC - Υβριδικά Νευρωνικά Δίκτυα

Η επιλογή των κατάλληλων χαρακτηριστικών που αντιπροσωπεύουν τα δεδομένα μας, μπορεί να χαρακτηριστεί ως μία δύσκολη αλλά σημαντική διαδικασία για την κατασκευή ενός αποδοτικού ταξινομητή στην αναγνώριση ακουστικών γεγονότων. Καθώς τα δεδομένα μας έχουν την μορφή ακουστικών αρχείων, τα χαρακτηριστικά που χρησιμοποιήθηκαν για την εκπαίδευση HMM και DNN ταξινομητών είναι οι φασματικοί συντελεστές της κλίμακας Mel ή αλλιώς MFCC για συντομογραφία. Μετά από την τμηματοποίηση των αρχείων της βάσης, ακολουθήθηκε μία αυστηρή διαδικασία για την εξαγωγή των διανυσμάτων χαρακτηριστικών. Θα αποτελούσε σοβαρή παράλειψη να μην αναφερθεί ότι τα χαρακτηριστικά MFCC εξάγονται σταδιακά από ένα ηχητικό σήμα με τη χρήση ενός κυλιόμενου παραθύρου. Περισσότερες πληροφορίες σχετικά με αυτή τη διαδικασία, θα παρουσιαστούν στα ακόλουθα υποκεφάλαια. Ύστερα από τη μελέτη του τύπου χαρακτηριστικών, ακολουθεί εκτενής ανάλυση των δύο τύπων ταξινομητών που χρησιμοποιήθηκαν στο πλαίσιο της διπλωματικής, των κρυφών μοντέλων Markov και των βαθιών νευρωνικών δικτύων καθώς και την προσαρμογή του προβλήματός μας σε αυτούς.

2.1 Φασματικοί συντελεστές κλίμακας Mel

Το ανθρώπινο αυτί αντιλαμβάνεται πιο εύκολα τις αλλαγές σε έναν ήχο στις χαμηλές συχνότητες παρά στις υψηλές. Η κλίμακα Mel σχεδιάστηκε ώστε τα διανύσματα χαρακτηριστικών να αντιπροσωπεύουν αυτή τη δυνατότητα. Στην αναγνώριση ακουστικών γεγονότων το επιθυμητό αποτέλεσμα είναι ο εντοπισμός και η αναγνώριση των γεγονότων που μπορεί να ακούσει ο άνθρωπος, αποκλείοντας άλλες συχνότητες. Η μετατροπή της συχνότητας από την κλίμακα Hertz στην κλίμακα Mel πραγματοποιείται με τον ακόλουθο μαθηματικό τύπο:

$$M(f) = 1125 \ln(1 + \frac{f}{700}) \tag{2.1}$$

Η αντίστροφη μετατροπή γίνεται ως εξής:

$$M^{-1}(m) = 700(\exp \frac{m}{1125} - 1) \tag{2.2}$$

όπου f και m είναι η συχνότητα στην κλίμακα των Hertz και Mel αντίστοιχα.

Οι συντελεστές MFCC προτάθηκαν τη δεκαετία του ′80 από τους S. Davis και P. Mermelstein [11] και έκτοτε χρησιμοποιούνται σαν χαρακτηριστικά που περιγράφουν ηχητικά σήματα στην ηχητική και φωνητική αναγνώριση. Τα βήματα που πρέπει να ακολουθηθούν για τον υπολογισμό των συντελεστών Mel παρουσιάζονται παρακάτω και στο Σχήμα 2.1:

1. Ορίζεται ένα παράθυρο-πλαισίου (προκαθορισμένη τιμή τα 25ms). Το ηχητικό σήμα είναι ορισμένο στον χρόνο και μεταβάλλεται σε συγκεκριμένες χρονικές στιγμές. Σε πληθώρα ηχητικών σημάτων, τέτοιες αλλαγές συμβαίνουν κάθε ms. Έχοντας αυτό κατά νου, καθίσταται πλήρως κατανοητό πως τα διανύσματα χαρακτηριστικών θα εξαχθούν σε αυτά τα χρονικά διαστήματα (τις χρονικές περιόδους που το σήμα δεν αλλάζει). Έτσι λοιπόν το σήμα τμηματοποιείται σε 25ms πλαίσια. Το πλαίσιο που ορίζουμε πρέπει να έχει χρονική διάστημα 20-40ms, αφού μικρότερες τιμές από 20 θα μας παρέχουν ανεπαρκή πληροφορία, ενώ τιμές μεγαλύτερες του 40 θα αντιπροσωπεύουν πολλαπλές αλλαγές στο σήμα.

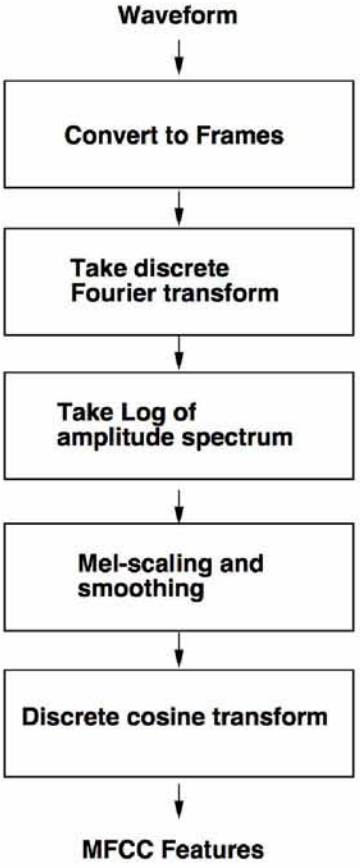
2. Για κάθε πλαίσιο υπολογίζεται ο διακριτός μετασχηματισμός Fourier (discrete Fourier transform ή DFT).

3. Εφαρμόζεται το φίλτρο Mel στο φάσμα ισχύος και αθροίζεται η ενέργεια σε κάθε πλαίσιο. Συγκεκριμένα, το πρώτο φίλτρο υποδεικνύει την ενέργεια κοντά σε μηδενικές συχνότητες. Η κλίμακα Mel καθορίζει πόσο ευρύ θα είναι κάθε φίλτρο για να υπολογιστεί η ενέργεια.

4. Υπολογίζεται ο λογάριθμος όλων των ενεργειών με σκοπό τα χαρακτηριστικά που θα εξαχθούν να προσεγγίζουν τι ακούει το ανθρώπινο αυτί.

5. Εφαρμόζεται ο διακριτός μετασχηματισμός συνημιτόνου (discrete cosine transform ή DCT) των προηγούμενων αποτελεσμάτων. Αυτή η ενέργεια αποσκοπεί στο να εξαλήψει της επικαλύψεις-συσχετίσεις των φίλτρων. Ο DCT χρησιμοποιείται σαν μέσο συμπίεσης της πληροφορίας του σήματος στους συντελεστές Mel και αποσυσχετίζει τις ενέργειες, πράγμα που επιτρέπει την κατασκευή διαγώνιων πινάκων συνδιασποράς (χρήσιμο στην εκπαίδευση HMM) [12].

6. Σαν διάνυσμα χαρακτηριστικών κρατάμε τους συντελεστές 1 έως 13 και απορρίπτονται οι υπόλοιποι. Αυτό συμβαίνει γιατί υψηλότεροι συντελεστές DCT συνεπάγονται γρήγορες αλλαγές στην ενέργεια του σήματος, άρα και του ίδιου ηχητικού σήματος, πράγμα που υποβαθμίζει την απόδοση συστημάτων φωνητικής και ηχητικής αναγνώρισης.



Σχημα 2.1: Ροή εργασίας για την εξαγωγή των χαρακτηριστικών. Λήφθηκε από το [3].

Στο συγκεκριμένο σχέδιο εργασίας χρησιμοποιήσαμε τους συντελεστές Delta-Delta MFCC, γνωστοί και ως διαφορικοί συντελεστές. Το διάνυσμα χαρακτηριστικών MFCC περιγράφει μόνο το φάσμα ισχύος ενός πλαισίου, αλλά φαίνεται ότι τα ηχητικά γεγονότα θα έχουν επιπλέον πληροφορία στη δυναμική τους. Αποδεδειγμένα ο υπολογισμός των τροχιών MFCC και η προσάρτησή τους στο αρχικό διάνυσμα χαρακτηριστικών αυξάνει την απόδοση της αναγνώρισης. Συγκεκριμένα, αν έχουμε 13 συντελεστές MFC, θα προσθέσουμε 13 συντελεστές delta, βάσει του τύπου 2.3, οι οποίοι θα συνδυάζονταν για να δώσουμε ένα διάνυσμα με 26 χαρακτηριστικά.

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{2.3}$$

όπου το d_t είναι ένας συντελεστής delta από το πλαίσιο t και υπολογισμένος από τους στατικούς όρους c_{t+n} έως c_{t-n}. Τυπική τιμή για το N είναι 2. Οι συντελεστές delta-delta(επιτάχυνσης) υπολογίζονται με τον ίδιο ακριβώς τρόπο, αλλά στη θέση των στατικών όρων τοποθετούμε τους συντελεστές delta.

2.2 Κρυφά Μοντέλα Markov

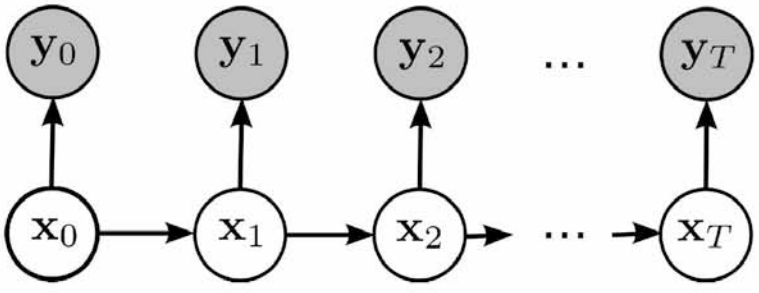
Τα κρυφά μοντέλα Markov αποτελούν ένα εργαλείο για την μοντελοποίηση δεδομένων που έχουν υπόσταση στον χρόνο. Τα HMM χρησιμοποιούνται κυρίως στην ηχητική και φωνητική αναγνώριση και καθιστούν την αναπαράσταση της κατανομής πιθανότητας σε μία ακολουθία παρατηρήσεων. Θεωρώντας ότι μία παρατήρηση συμβολίζεται με Y και η κατάσταση με S, τότε η παρατήρηση και η κατάσταση τη χρονική στιγμή t μπορεί να συμβολιστεί Y_t και S_t αντίστοιχα. Τα κρυφά μοντέλα Markov διατηρούν τις ιδιότητες των μοντέλων Markov, αλλά η λέξη-κλειδί *κρυφά* αναφέρεται σε δύο βασικές ιδιότητες:

1. Μία παρατήρηση τη χρονική στιγμή t παρήχθη στην κατάσταση S_t που είναι κρυφή από τον παρατηρητή.
2. Η κάθε κατάσταση ακολουθεί τις ιδιότητες Markov, το οποίο σημαίνει πως οι τιμές των μεταβλητών στην κατάσταση S_t εξαρτώνται μόνο από την κατάσταση S_{t-1} και όχι από τις $t-2$ προηγούμενες.

Ο τύπος για την από κοινού κατανομή πιθανότητας της ακολουθίας των καταστάσεων είναι ο εξής:

$$P(S_{1:T}, Y_{1:T}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(Y_T|S_T) \tag{2.4}$$

Το μοντέλο πιθανότητας HMM παρουσιάζεται στο Σχήμα 2.2



Σχημα 2.2: Παράδειγμα κρυφού μοντέλου Markov. Λήφθηκε από το [4].

2.3 Βαθιά νευρωνικά δίκτυα

2.3.1 Αρχιτεκτονικές βαθιάς μάθησης

Κατά την υλοποίηση αυτής τη μεταπτυχιακής εργασίας, πειραματιστήκαμε με την αρχιτεκτονική της βαθιάς μάθησης και συγκεκριμένα τα νευρωνικά δίκτυα. Ένα βαθύ νευρωνικό δίκτυο αποτελεί ένα τεχνητό δίκτυο με πολλαπλά κρυφά επίπεδα μεταξύ της εισόδου και της εξόδου. Τα βαθιά νευρωνικά δίκτυα ή DNN εκπαιδεύονται με τη χρήση του αλγορίθμου οπισθοδιάδοσης. Ύστερα από κάθε επανάληψη, τα βάρη ανανεώνονται μέσα από τον ακόλουθο τύπο:

$$w(t+1) = w(t) + \eta \Delta w \tag{2.5}$$

όπου η είναι ο ρυθμός εκπαίδευσης, $w(t)$ η τρέχουσα εκτίμηση των βαρών και Δw η αντίστοιχη διόρθωση ώστε να προκύψει η επόμενη εκτίμηση $w(t+1)$. Η επιλογή της συνάρτησης κόστους έγκειται αποκλειστικά στον τύπο των δεδομένων (με επίβλεψη, χωρίς επίβλεψη) και στην συνάρτηση ενεργοποίησης (βηματική, σιγμοειδής). Για παράδειγμα, στο HTK χρησιμοποιήθηκε η συνάρτηση διεντροπίας (cross-entropy) σαν συνάρτηση κόστους:

$$J_{ce} = -\sum_{i=1}^{N}\sum_{k=1}^{k_L} y_k(i) \ln \frac{\hat{y}_k(i)}{y_k(i)} \tag{2.6}$$

και σαν συνάρτηση εξόδου του DNN η softmax:

$$\hat{y}_k = \frac{\exp(v_k^L)}{\sum_k \exp(v_{k'}^L)} \tag{2.7}$$

όπου N είναι το πλήθος των διανυσμάτων εκπαίδευσης, L το πλήθος των κρυφών επιπέδων και v η συνάρτηση ενεργοποίησης. Τέλος, οι μεταβλητές y_k και $\hat{y}_k$ δηλώνουν την επιθυμητή και πραγματική αντίστοιχα εξόδο του νευρωνικού δικτύου. Εκτός από την παραπάνω έχουμε και την συνάρτηση μέσου σφάλματος τετραγώνων (mean square error ή MSE), η οποία δίνεται από τον τύπο:

$$J_{mse} = -\frac{1}{k_L}\sum_{i=1}^{N}\sum_{k=1}^{k_L}(y_k(i) - \hat{y}_k(i))^2 \tag{2.8}$$

Οι τιμές και των δύο συναρτήσεων κόστους είναι θετικές και κριτήριό τους είναι να ελαχιστοποιήσουν την συνάρτηση ζεύγος τους. Συγκεκριμένα, η συνάρτηση διεντροπίας χρησιμοποιείται με τη συνάρτηση εξόδου Softmax. Όταν λοιπόν έχουμε τέτοια ζεύγη συναρτήσεων, η παράγωγος της συνάρτησης J σε σχέση με την τιμή της συνάρτησης εξόδου

θα είναι:

$$\frac{\partial J}{\partial J_o ut(i)} = y_k(i) - \hat{y}_k(i) \tag{2.9}$$

Το οποίο εξυπηρετεί στον υπολογισμό της διαφορά μεταξύ της εκτιμώμενης και της πραγματικής τιμής εξόδου που βασίζεται ο αλγόριθμος της οπισθοδιάδοσης[13].

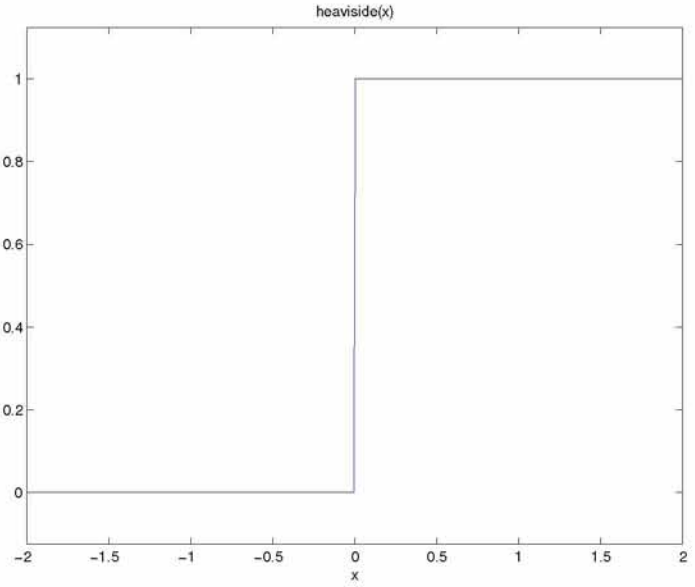
2.3.2 Συνάρτηση ενεργοποίησης

Στις παραπάνω περιπτώσεις τα κρυφά επίπεδα δεν ήταν παρόντα. Αν σε ένα νευρωνικό δίκτυο υπάρχουν κρυφά επίπεδα μεταξύ εισόδου και εξόδου, κάθε νευρώνας του $i-$στου κρυφού επιπέδου θα είναι 0 ή 1, το οποίο σημαίνει ότι κάλλιστα η συνάρτηση βήματος μπορεί να χρησιμοποιηθεί για αυτή την εναλλαγή τιμών. Η συμπεριφορά της συνάρτησης βήματος παρουσιάζεται στο Σχήμα 2.3. Για να αποφευχθεί όμως αυτή η απότομη μετάβαση από το 0 στο 1, χρησιμοποιήθηκε η σιγμοειδής συνάρτηση που παρουσιάζεται στο Σχήμα 2.4. Οι μαθηματικοί τύποι που περιγράφουν τις τρεις βασικές συναρτήσεις ενεργοποίησης είναι:

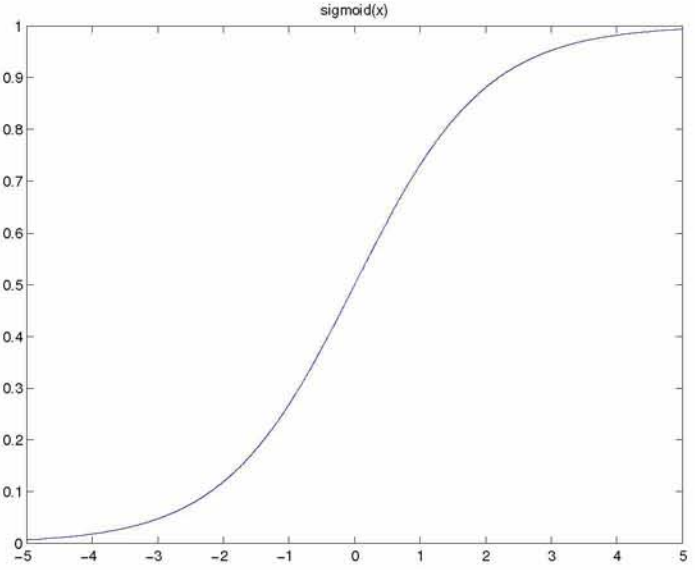
$$step(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases} \tag{2.10}$$

$$sigm(x) = \frac{1}{1 + e^{-\alpha x}} \tag{2.11}$$

$$tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \tag{2.12}$$

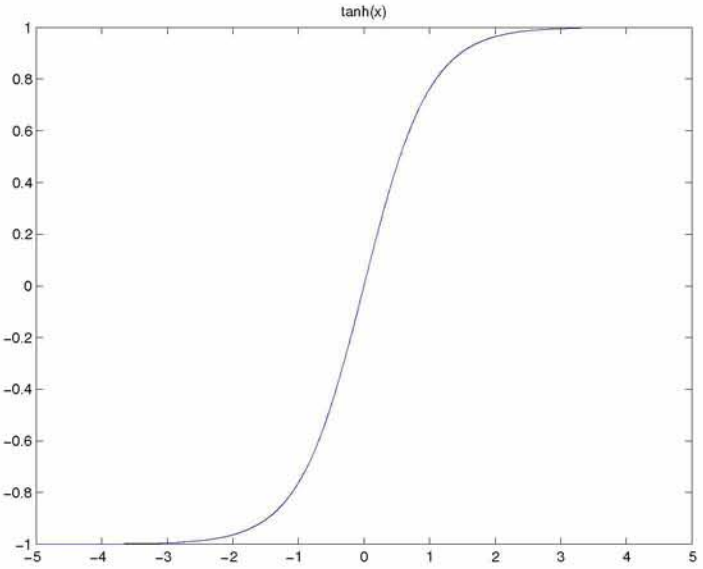


Σχημα 2.3: Συνάρτηση βήματος



Σχημα 2.4: Σιγμοειδής συνάρτηση

Σε ένα δηλαδή νευρωνικό δίκτυο με δύο κρυφά επίπεδα, η είσοδος στους νευρώνες του 2ου επιπέδου θα έχουν τιμές που ανήκουν στο διάστημα $[0, 1]$.



Σχημα 2.5: Συνάρτηση υπερβολικής εφαπτομένης

Η συνάρτηση υπερβολικής εφαπτομένης (hyperbolic tangent function tanh) μετατοπίζει την έξοδο στο $[-1, 1]$.

2.4 Υβριδικό μοντέλο μάθησης

2.4.1 Η θεωρία

Τα υβριδικά νευρωνικά δίκτυα αποτελούν μία ακόμη προσέγγιση στη μεγάλη μελέτη που αφορά τις αρχιτεκτονικές βαθιάς μάθησης. Ειδικότερα, τα συστήματα αυτά, που ενσωματώνουν τα βαθιά νευρωνικά δίκτυά και τα κρυφά μοντέλα Markov, έχουν επιτύχει πρόσφατα αξιοσημείωτη απόδοση σε θέματα αναγνώρισης ομιλίας με μεγάλα λεξιλόγια. Τα συστήματα αυτά, ωστόσο, εξακολουθούν να βασίζονται στα κρυφά μοντέλα Markov και υπολογίζουν τις ακουστικές βαθμολογίες για το παραθυρωμένα πλαίσια, ανεξάρτητα το ένα από το άλλο, πάσχοντας έτσι από την ίδια δυσκολία όπως στα συστήματα GMM-HMM[14].

Η γενική ιδέα έχει ως εξής:

- σε ένα μοντέλο HMM, αντικαθιστούμε τα γκασουσιανά μοντέλα μίξης, που υπολόγιζαν τις συναρτήσεις πυκνότητας πιθανότητας, με τις εξόδους των νευρωνικών δικτύων.
- μετατρέπουμε τις εκτιμήσεις των εκ των υστέρων πιθανοτήτων σε κλίμακα πιθανοτήτων (scaled likelihood), διαιρώντας με τις σχετικές συχνότητες των δεδομένων

εκπαίδευσης για κάθε κλάση.

$$P(x_t|C_k) \propto \frac{P(C_k|x_t)}{P_{train}C_k} = \frac{y_k}{P_{train}C_k} \tag{2.13}$$

- πλέον οι έξοδοι του νευρωνικού δικτύου αντιστοιχούν στις κλάσεις των φωνημάτων. Στο συγκεκριμένο σχέδιο εργασίας, το μέγεθος της εξόδου ήταν 13 κλάσεις.

2.4.2 DNN-HMM εναντίον GMM-HMM

Τα πλεονεκτήματα του προαναφερθέντος μοντέλου είναι:

- Ενσωμάτωση πολλαπλών πλαισίων δεδομένων στην είσοδο.
- Είναι πιο ευέλικτο από το μοντέλο των GMM, καθώς το δεύτερο δεν είναι αποδοτικό για μη-γραμμικές κλάσεις.

Εκτός όμως από τα παραπάνω τα DNN-HMM μοντέλα χαρακτηρίζονται από τα ακόλουθα μειονεκτήματα[15]:

- Τα μονοφωνικά μοντέλα είναι ανεξάρτητα από το περιβάλλον.
- Αδύναμοι αλγόριθμοι προσαρμογής ομιλητών.
- Υπολογιστικά κοστοβόρο - πιο δύσκολο να παραλληλιστεί από τα γκαουσιανά μοντέλα μίξης.
- Είναι δύσκολο να δημιουργηθούν περίπλοκα μοντέλα όπως στα γκαουσιανά μοντέλα μίξης.

2.5 Προβλήματα με τα βαθιά νευρωνικά δίκτυα

Τα βαθιά νευρωνικά δίκτυα χαρακτηρίζονται από δύο βασικά προβλήματα. Το πρώτο είναι αυτό της υπερ-εκπαίδευσης και το δεύτερο είναι ο χρόνος υπολογισμού τους. Κατά τη διάρκεια εκπαίδευσης ενός DNN με το HTK, ακόμα και με τετραπύρηνο επεξεργαστή, ο χρόνος ολοκλήρωσης ήταν 9 ώρες. Μεγάλες βελτιώσεις έχουν γίνει όμως με τη χρήση της τεχνολογίας πολυνηματισμού και CUDA στις κάρτες γραφικών. Καθώς τα εργαλεία-συναρτήσεις έχουν υλοποιηθεί για να τρέχουν σε πολυπύρηνους επεξεργαστές, χρησιμοποιήσαμε τις αντίστοιχες βιβλιοθήκες για αυτή την εργασία. Σε αυτό το σημείο πρέπει να αναφερθεί πως η καθυστέρηση της διάτρεξης των νευρωνικών δικτύων δεν οφείλεται στο μέγεθος της βάσης δεδομένων, αλλά στις επιπλέον επαναλήψεις που λαμβάνουν χώρα κατά τη διαδικασία της

εκπαίδευσης, δημιουργώντας έτσι ανεπιθύμητες εξαρτήσεις.

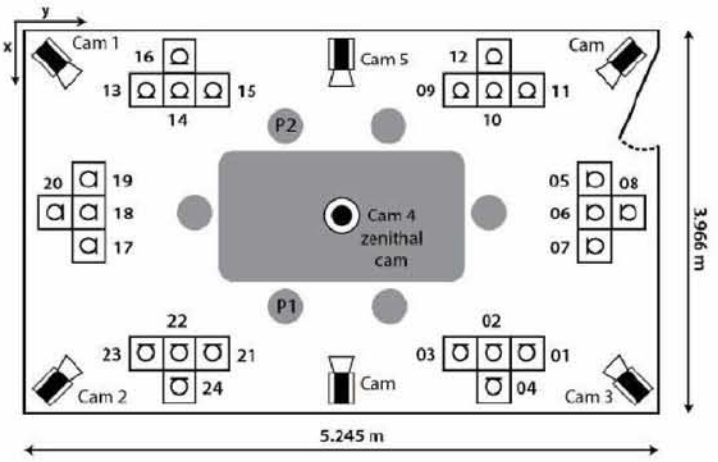
Ο αλγόριθμος οπισθοδιάδοσης σε συνδυασμό με αυτόν της απότομης κατάδυσης [16] (gradient descent) αποτελεί τον προτεινόμενο τρόπο εκπαίδευσης των νευρωνικών δικτύων, αν και το κόστος όπως αναφέρθηκε είναι ένα μειονέκτημα. Ωστόσο, τα βαθιά νευρωνικά δίκτυα τείνουν να είναι πιο αποδοτικά χάρη στην λιγότερη ανάγκη για παραμέτρους και υπολογιστικά στοιχεία, χωρίς όμως αυτό να σημαίνει ότι αποτελούν πάντοτε καλύτερη λύση από τις ρηχές μεθόδους. Η εκπαίδευση ενός βαθιού νευρωνικού δικτύου είναι γνωστό ότι αποτελεί μία δύσκολη διαδικασία. Κατά τον προκαθορισμένο τρόπο εκπαίδευσης τα βάρη του δικτύου αρχικοποιούνται με τυχαίες τιμές και εφαρμόζεται ο αλγόριθμος απότομης κατάδυσης, πράγμα που δίνει φτωχές λύσεις για νευρωνικά δίκτυα με 3 ή περισσότερα κρυφά επίπεδα. Για αυτό τον λόγο, τα νευρωνικά δίκτυα περιορίζονται στο ένα ή δύο κρυφά επίπεδα [17]. Η δήλωση αυτή θα γίνει πιο κατανοητή κατά την διαδικασία αξιολόγησης του ταξινομητή στο Κεφάλαιο 5 όπου παρουσιάζονται αναλυτικά τα αποτελέσματα.

Κεφάλαιο 3

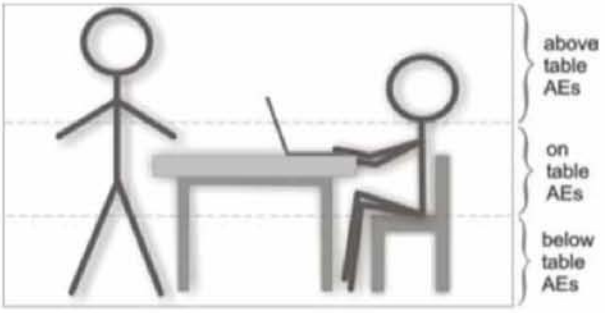
Η Πολυ-καναλική Βάση Δεδομένων UPC-TALP

3.1 Βασικές Πληροφορίες

Η συγκεκριμένη βάση δεδομένων καταγράφηκε κατά τη διάρκεια του CHIL Project (Computers in the Human Interaction Loop) στο πλαίσιο του ολοκληρωμένου προγράμματος εργασίας European Commission's Sixth Framework και στην ουσία περιέχει μία σειρά από ακουστικά αρχεία με γεγονότα που μπορούν να συμβούν σε ένα δωμάτιο συνεδριάσεων [18]. Τα ηχητικά αρχεία της βάσης ηχογραφήθηκαν από 24 συνολικά μικρόφωνα, τα οποία ήταν συγχρονισμένα και οργανωμένα σε ομάδες. Μέσα στην αίθουσα συνεδριάσεων υπήρχαν 6 τέτοιες T-σχήματος ομάδες μικροφώνων, όπως φαίνεται στο Σχήμα 3.1 και 3.2 όπου παρουσιάζεται η κάτοψη του δωματίου που έγιναν οι ηχογραφήσεις και ο τύπος των δεδομένων που προσπαθήσαμε να ταξινομήσουμε αντίστοιχα. Η επιπρόσθετη πληροφορία πάνω στο σχήμα είναι η θέση των δύο συμμετόχοντων P1 και P2. Μεταξύ των διαδοχικών γεγονότων υπήρχε παύση μερικών δευτερολέπτων, ώστε ο ταξινομητής που θα εκπαιδευτεί να αναγνωρίζει και την απουσία κάποιου ηχητικού γεγονότος (ησυχία). Εκτός της ηχητικής κάλυψης, το δωμάτιο ήταν εξοπλισμένο με έξι κάμερες που κατέγραφαν την όλη διαδικασία. Η συγκεκριμένη βάση δεδομένων μπορεί να χρησιμοποιηθεί σαν υλικό εκπαίδευσης για τεχνολογίες αναγνώρισης ακουστικών γεγονότων, όπως και για αλγορίθμους αποκωδικοποίησης, οι οποίοι δοκιμάζονται σε ήσυχα περιβάλλοντα χωρίς μόνιμες επικαλύψεις μεταξύ των ήχων.



Σχημα 3.1: Κάτοψη του UPC-δωματίου. Παρουσιάζονται οι θέσεις και η κατανομή των 24 μικροφώνων. Η εικόνα λήφθηκε από το [5].



Σχημα 3.2: Διαχωρισμός των δεδομένων ανάλογα με τον χώρο του δωματίου. Η εικόνα λήφθηκε από το [5].

3.2 Τεχνικές Πληροφορίες

Όπως προαναφέρθηκε, η βάση δεδομένων UPC-TALP περιέχει μεμονωμένα και αυθόρμητα ηχητικά συμβάντα σε ένα έξυπνο περιβάλλον. Η δομή της αποθηκευμένης βάσης παρουσιάζεται παρακάτω:

- 1^{st} DVD: S01, S02
- 2^{nd} DVD: S03, S04
- 3^{rd} DVD: S05, S06
- 4^{th} DVD: S07, T02, T03, T04
- 5^{th} DVD: S08, T01
- 6^{th} DVD: T05, T06, T07, T08, T09

Τα ακουστικά γεγονότα σε κάθε συνεδρία (S01-S08) παρήχθηκαν από έξι διαφορετικά άτομα. Σε κάθε προσπάθειά τους, στόχος ήταν να αναπαράγουν μία συγκεκριμένη ακολουθία από ήχους που θα καταγράφονταν από τα 24 μικρόφωνα. Πιο ειδικά, η συχνότητα στην οποία πραγματοποιήθηκαν οι ηχογραφήσεις ήταν τα 44.1kHz και τα αρχεία αποθηκεύτηκαν σε μορφή *.wv. Στη βάση δεδομένων έχουν κατηγοριοποιηθεί 12 κλάσεις ηχητικών γεγονότων που μπορούν να συμβούν ρεαλιστικά σε ένα δωμάτιο συνεδριάσεων. Από αυτά, το χειροκρότημα και το γέλιο συνέβησαν όταν στο δωμάτιο ήταν παραπάνω από ένας συμμετέχοντες. Το πλήθος των γεγονότων όπως και η κατανομή τους κατά το πέρας της καταγραφής της βάσης δεδομένων παρουσιάζεται στον Πίνακα 3.1.

Ηχητικό Γεγονός	S01	S02	S03	S04	S05	S06	S07	S08
Χτύπημα Πόρτας	9	8	10	10	10	8	11	13
Κλείσιμο Πόρτας	17	15	19	20	40	37	56	52
Βήματα	10	10	8	23	43	34	28	50
Μετακίνηση Καρέκλας	19	37	32	22	23	38	34	40
Κουτάλι/Κουδούνισμα φλιτζανιού	10	11	13	11	10	15	11	15
Τύλιγμα Χαρτιού Εργασίας	9	11	10	8	17	12	12	12
Ήχος Κλειδιών	11	11	11	8	0	13	10	18
Ήχος Πληκτρολογίου	10	10	13	12	10	13	10	11
Ήχος Τηλεφώνου/Μουσική	11	18	11	14	8	11	13	15
Χειροκρότημα	9	5	9	11	12	9	14	14
Βήχας	10	10	12	13	9	13	11	12
Ομιλία	0	0	0	0	8	20	12	34

ΠΙΝΑΚΑΣ 3.1: Κατανομή των κλάσεων της βάσης δεδομένων UPC-TALP ανά συνεδρία.

Τα σύμβολα των κλάσεων (ηχητικών γεγονότων) μαζί με τη διάρκεια τους καταγράφονται ανά συνεδρία σε ένα *.csv αρχείο.

Κεφάλαιο 4

Ροή εργασίας & πειραματικά Εργαλεία

4.1 HTK

Το HTK (εργαλειοθήκη κρυφών μοντέλων Markov) αναπτύχθηκε στο εργαστήριο Μηχανικής Μάθησης του Τμήματος Μηχανικών στο Πανεπιστήμιο του Cambridge με σκοπό τη δημιουργία και τη διαχείριση των μοντέλων Markov [13]. Το σύνολο του εργαλείου έχει υλοποιηθεί σε γλώσσα C, ομοίως και τα εκτελέσιμα αρχεία και οι βιβλιοθήκες του. Τα προαναφερθέντα είναι χωρισμένα σε φακέλους για την ευκολότερη εύρεση από τον χρήστη. Σχετικά με την εγκατάσταση, αρκεί η μεταγλώττιση και παραγωγή του εκτελέσιμου κώδικα C (περισσότερες πληροφορίες εδώ http://htk.eng.cam.ac.uk). Παρόλο που το HTK εξειδικεύεται σε έρευνες σχετικά με την φωνητική αναγνώριση, έχει πολυάριθμες εφαρμογές συμπεριλαμβανομένων της φωνητικής σύνθεσης και αναγνώρισης γραμματικών χαρακτήρων. Μέσα από το HTK, ο χρήστης δημιουργεί HMM μοντέλα, επεξεργάζεται τις ιδιότητες των Γκαουσιανών κατανομών, τα εκπαιδεύει με δικές του παραμετρικοποιήσεις και τέλος χρησιμοποιεί τον αλγόριθμο Viterbi για να αναγνωρίσει τα ηχητικά γεγονότα. Στην τελευταία του έκδοση 3.5 (2015 beta) δίνεται η δυνατότητα στον χρήστη να δημιουργήσει μοντέλα βαθιών αρχιτεκτονικών όπως τα νευρωνικά δίκτυα. Η χρήση του βασίζεται σε script αρχεία που διευκολύνουν τη διαδικασία της εκπαίδευσης και κατηγοριοποίησης της βάσης δεδομένων.

4.1.1 Προετοιμασία δεδομένων

Σε αυτό το στάδιο του σχεδίου εργασίας τα ακουστικά αρχεία της βάσης δεδομένων έχουν επεξεργαστεί και χωριστεί σε ξεχωριστούς φακέλους ανάλογα με τη συνεδρία στην οποία ανήκουν. Έτσι, δημιουργήθηκε ο φάκελος **train** με δύο υπο-φακέλους **mfcc** και **wav** που περιέχουν τα διανύσματα χαρακτηριστικών και τα ακουστικά αρχεία αντίστοιχα. Αρχικά, τα MFCC χαρακτηριστικά εξήχθησαν από κάθε ακουστικό αρχείο και τοποθετήθηκαν

(αυτόματα πάντα) στο κατάλληλο φάκελο. Γι΄ αυτό τον λόγο δημιουργήθηκε ένα αρχείο ρυθμίσεων με το όνομα **config**. Το περιεχόμενο αυτού παρουσιάζεται παρακάτω στο Σχήμα 4.2.

```
SOURCEFORMAT = WAV
TARGETKIND = MFCC_0_D_A
TARGETRATE = 100000.0
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 12
```

Σχημα 4.1: Αρχείο ρυθμίσεων

Το παραπάνω περιεχόμενο ουσιαστικά περιγράφει τις ιδιότητες των χαρακτηριστικών με τις οποίες θα εξαχθούν τα χαρακτηριστικά MFCC. Κρατώντας τις πιο σημαντικές από αυτές, έχουμε:

- **SOURCEFORMAT**, δηλώνει τον τύπο του ακουστικού αρχείου εισόδου, WAV στην περίπτωσή μας.
- **TARGETKIND MFCC_0_D_A**, εξάγει τα διανύσματα MFCC διάστασης 39 αντικαθιστώντας τον τελευταίο με τον μηδενικό συντελεστή που αντιστοιχεί στην ενέργεια του κάθε παραθύρου. Η παράμετρος αυτή απαιτεί μία σειμβολοσειρά που προσδιορίζει αυτή την ιδιότητα. Με σκοπό τη βελτίωση του συνόλου εκπαίδευσης χρησιμοποιήθηκαν και οι 39 συντελεστές MFCC. Έτσι η παράμετρος TARGETKIND άλλαξε σε MFCC_0_D_A, το οποίο πρόσθεσε τους delta και delta delta συντελεστές.
- **WINDOWSIZE**, το μέγεθος του πλαισίου με το οποίο σαρώνουμε το ηχητικό σήμα. Η τιμή του είναι προκαθορισμένη στα 25ms και απαιτεί μία θετική πραγματική τιμή.
- **TARGETRATE**, δηλώνει κατά πόσο κινούμαστε πάνω στο σήμα. Για παράδειγμα, το αρχικό παράθυρο διαρκεί από τη χρονική στιγμή 0 μέχρι 25ms και το επόμενο από 10 έως 35ms.
- **NUMCHANS**, απαιτεί έναν ακέραιο αριθμό που δηλώνει τον αριθμό των filter-bank channels που χρησιμοποιούνται στην ανάλυση.

- **NUMCEPS**, αποτελεί τη μεταβλητή που υποχρεώνει την διατήρηση των 13 πρώτων συντελεστών. Από τη στιγμή που έχουμε δηλώσει το TARGETKIND για Delta Delta MFCC, θα δεσμευτούν οι 39 συντελεστές.

Με βάση την παραπάνω οδηγία, δημιουργήσαμε τα ακόλουθα σύνολα αρχείων MFCC:

- μεμονωμένα αρχεία εκπαίδευσης από τις συνεδρίες 1 έως 7.
- μεμονωμένα αρχεία για αποκωδικοποίηση από τη συνεδρία 8.
- ενσωματωμένα αρχεία εκπαίδευσης (ένα αρχείο χαρακτηριστικών για κάθε μικρόφωνο) από τις συνεδρίες 1 έως 7.
- ενσωματωμένα αρχεία για αποκωδικοποίηση από τη συνεδρία 8.

Τέλος, τοποθετήθηκαν αρχεία ετικετών στα ακουστικά αρχεία. Μέσα από τα αρχεία ετικετών δηλώνεται σε ποιο συμβάν αντιστοιχεί το αρχείου ήχου ή χαρακτηριστικών. Για παράδειγμα, το τμήμα ενός αρχείου ετικετών έχει τη μορφή του Σχήματος 4.2.

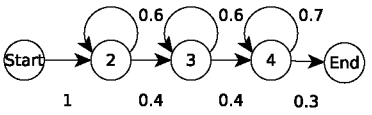
```
#!MLF!#
"*/S01-eventid_0.lab"
kn
.   "*/S01-eventid_1.lab"
si
.
"*/S01-eventid_2.lab"
ds
.
"*/S01-eventid_3.lab"
si
.
"*/S01-eventid_4.lab"
ds
.
"*/S01-eventid_5.lab"
si
```

Σχημα 4.2: Το αρχείο αποτελεί ένα Κύριο Αρχείο Ετικετών Master Label File (MLF).

4.1.2 Δημιουργία ευθυγράμμισης κατάστασης-παραθύρου

Το πρώτο βήμα στη δημιουργία του υβριδικού νευρωνικού δικτύου είναι η παραγωγή των ευθυγραμμίσεων κατάστασης-παραθύρου στα δεδομένα εκπαίδευσης. Με άλλα λόγια

αντιστοιχούμε το κάθε παράθυρο-γεγονός στην κατάσταση του HMM από όπου πήραμε την πιθανότητά του. Στη συγκεκριμένη περίπτωση, οι καταστάσεις που εκπέμπουν πιθανότητες, είναι η 2, 3 και 4, όπως παρουσιάζεται στο παρακάτω σχήμα.



Σχημα 4.3: Πρωτότυπο μοντέλο HMM.

Βασική προυπόθεση για τα παραπάνω είναι η ύπαρξη ενός προ-εκπαιδευμένου μοντέλου με τα φωνήματα. Όπως αναφέρεται και στη βιβλιογραφία [13], όσο καλύτερο είναι το προ-εκπαιδευμένο μοντέλο, τόσο καλύτερες θα είναι οι ευθυγραμμίσεις καταστάσεων που οδηγούν σε ένα υβριδικό μοντέλο με χαμηλό ποσοστό λάθους. Η ευθυγράμμιση πραγματοποιείται με την κλήση της εντολής του HTK, HVite:

HVite -C config.align -H pretrainedModel -S train.scp -I train.ref.mlf -i train.aligned.mlf -f -o MW -b sil -a -y lab dictionary modelList

- **όπου, *pretrainedModel***, είναι ο ορισμός του προ-εκπαιδευμένου μοντέλου GMM-HMM.
- ***modelList***, περιέχει το σύνολο των κλάσεων.
- ***config.align***, περιέχει τις βασικές ρυθμίσεις του προεκπαιδευμένου μοντέλου

Η έξοδος της εντολής μας δίνει αρχεία του ακόλουθου τύπου:

```
 #!MLF!#
"S0001.lab"
0 100000 sil[2] -31.289440
100000 200000 sil[3] -50.577595
200000 300000 sil[4] -52.500790
300000 400000 sil[2] -51.115330
400000 500000 sil[3] -49.983665
500000 800000 sil[4] -150.454163
800000 2000000 sil-w+ah[2] -735.791443
2000000 2500000 sil-w+ah[3] -125.901878
2500000 2800000 sil-w+ah[4] -124.215851
```

Σχημα 4.4: Αρχείο εξόδου από κλήση της HVite

4.1.3 Κατασκευή προτύπου DNN-HMM

Πρώτο βήμα σε αυτό το στάδιο είναι να ορίσουμε ένα αρχικό μοντέλο με μηδενικές παραμέτρους σε όλα τα επίπεδα του νευρωνικού δικτύου. Σε αυτό το σημείο πρέπει να αναφερθεί ότι ορίσαμε δύο βασικά πρότυπα, ένα μεγέθους 39x128x39 και ένα 39x256x39. Αναλύοντας τις προαναφερθέντες τιμές έχουμε:

- ***39***, ισούται με το μέγεθος της εισόδου, δηλαδή των διανυσμάτων χαρακτηριστικών.
- ***128 ή 256***, είναι το μέγεθος του κρυφού επιπέδου (αριθμό κόμβων).
- ***39***, επειδή έχουμε 13 κλάσεις (μαζί με το γεγονός silence) πολλαπλασιασμένες με τις 3 εκπέμπουσες πιθανοτικές HMM καταστάσεις.

Όπως κάθε αρχείο του HTK, οι αρχικές γραμμές του αρχείου μοντελοποίησης του DNN-HMM περιγράφουν τον τύπο της εισόδου, που στην προκειμένη περίπτωση είναι τα χαρακτηριστικά Delta Delta MFCC. Οπότε έχουμε:

```
 o
<STREAMINFO> 1 39
<VECSIZE> 39<NULLD><MFCC_D_A_0><DIAGC>
M "layer2_weight"
<MATRIX> 128 39 (128x39 times)
V "layer2_bias"
<VECTOR> 128
0.000000e+00 (128 times)
M "layerout_weight"
<MATRIX> 39 500
V "layerout_bias"
<VECTOR> 39
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
```

Σχημα 4.5: Περιεχόμενο αρχείου dnn.initialModel

Μπορεί να παρατηρηθεί πως η διάσταση του μείγματος χαρακτηριστικών είναι ίση με τον αριθμό της στήλης του πίνακα βάρους, ο οποίος είναι ο αριθμός των μονάδων εισόδου αυτού

του επιπέδου και ο αριθμός γραμμών του πίνακα βάρους ταιριάζει με το μέγεθος των τιμών bias και είναι ίσος με τις μονάδες εξόδου. Έχοντας τα παραπάνω, το νευρωνικό δίκτυο έχει τον παρακάτω ορισμό.

```
 N "DNN"
<BEGINANN>
<NUMLAYERS> 3
<LAYER> 2
L "layer2"
<LAYER> 3
L "layerout"
<ENDANN>
```

Σχημα 4.6: Δομή πρότυπου νευρωνικού δικτύου

Χρησιμοποιώντας το πολυεπίπεδο perceptron, μπορούμε πλέον να συσχετίσουμε κάθε κόμβο εξόδου του νευρωνικού δικτύου με τις υπάρχουσες HMM καταστάσεις για να αρχικοποιηθούν οι παράμετροι του υβριδικού μοντέλου. Σε αυτό το σημείο χρησιμοποιήθηκε η εντολή HHEd του HTK.

HHEd -H hmm_mono/model -M dnn/init connect.hed class_list

όπου, **model** και **class_list** είναι οι μονοφωνητικές καταστάσεις HMM και η λίστα με τις κλάσεις αντίστοιχα. Το συνδεδεμένο μοντέλο, σαν έξοδος της εντολής, αποθηκεύεται στο φάκελο **init**. Τέλος, το αρχείο **connect.hed** περιέχει τις ακόλουθες απαραίτητες εντολές:

```
CH proto/dnn proto/foolist  N "DNN"
<HYBRID>
SW 1 39
SK MFCC_0_D_A
EL  L "layer2"
EL  L "layerout"
```

Σχημα 4.7: Δομή αρχείου connect.hed

Η εντολή CH συνδέει τις μονάδες εξόδου του νευρωνικού δικτύου με τις μονοφωνικές καταστάσεις HMM. Το αρχείο proto/dnn είναι το πρωτότυπο MLP μοντέλο που ορίσαμε και proto/foolist είναι ένα κενό αρχείο, αφού proto/dnn δεν περιείχε καταστάσεις HMM. Οι εντολές EL αρχικοποιούν τυχαία τα επίπεδα layer2 και layerout. Το μονοφωνικό μοντέλο DNN-HMM είναι η έξοδος.

4.1.4 Εκπαίδευση υβριδικού μοντέλου σε επίπεδο παραθύρου

Τα ακουστικά μοντέλα DNN εκπαιδεύονται χρησιμοποιώντας τα κριτήρια επιπέδου-πλασίου. Δεδομένου ότι το ορισμένο μοντέλο χρησιμοποιεί σαν συνάρτηση εξόδου τη softmax, η συνάρτηση κόστους που ταιριάζει είναι αυτή της διεντροπίας για την εκπαίδευση σε επίπεδο παραθύρου. Το στάδιο αυτό της εκπαίδευσης μπορεί να χωριστεί σε επιμέρους δύο βήματα, αυτά της προ-εκπαίδευσης (pre-training) και τέλος της βελτιστοποίησης παραμέτρων (fine-tuning).

Πριν τη χρήση της HNTrainSGD, για την εκπαίδευση του DNN-HMM μοντέλου, οι μέσοι όροι και οι διασπορές υπολογίστηκαν, καθώς όταν τα διανύσματα χαρακτηριστικών είναι κανονικοποιημένα σε μηδενική μέση τιμή και μοναδιαία διασπορά μπορούν να βελτιώσουν την ταχύτητα εκπαίδευσης. Το διάνυσμα διασποράς υπολογίζεται από τη συνάρτηση HCompV:

HCompV -p *%%% -k *.%%% -C cvn.cfg -q v -c cvn -S dnn.hv.scp

όπου, **dnn.hv.scp** αποτελεί ένα τυχαίο 10% του συνόλου εκπαίδευσης. Το υπόλοιπο αποθηκεύεται στο αρχείο **dnn.train.scp**.

4.1.4.1 Στάδιο pre-train

Βήμα 1

Έχοντας ορίσει τα παραπάνω, ξεκίνησε η εκπαίδευση κατά επίπεδο του υβριδικού μοντέλου. Τα επόμενα κρυφά επίπεδα προστέθηκαν σταδιακά μέχρι το δίκτυο να αποκτήσει το επιθυμητό μέγεθος. Πρέπει να ανεφερθεί πως όταν προστίθεται ένα νέο επίπεδο η HNTrainSGD χρησιμοποιείται για να εκπαιδευτεί το μοντέλο για μία διάτρεξη (epoch). Έχοντας σαν αναφορά το μοντέλο που ορίσαμε στην παράγραφο 4.1.3, εκπαιδεύσαμε το αρχικό υβριδικό δίκτυο καλώντας την:

HNTrainSGD -C config.basic -C config.pretrain -H dnn/init/model -M dnn/1layer -S dnn.train.scp -N dnn.hv.scp -l LABEL -I mono.aligned.mlf hmm_mono/monolist

όπου, το μοντέλο στην έξοδο αποθηκεύτηκε στον φάκελο **dnn/1layer**, οι ετικέτες εκπαίδευσης προήλθαν από το αρχείο **mono.aligned.mlf** και οι παράμετροι της εκπαίδευσης ήταν:

```
HANNET: MINIBATCHSIZE = 200
CRITERION = XENT
LEARNRATE = 0.001
MOMENTUM = 0.5
UPDATECLIP = 0.32
NMKLTHREADS = 4
```

Σχημα 4.8: Περιεχόμενο αρχείου config.pretrain.

Η τελευταία παράμετρος προστέθηκε εκ των υστέρων. Ειδικότερα, αφού μεταγλωττίσαμε το HTK με τις βιβλιοθήκες MKL της Intel με σκοπό την δυνανότητα πολυνηματισμού. Καθώς ο επεξεργαστής του Η/Υ είναι τετραπύρηνος με 8 νήματα, δώσαμε την ανάλογη τιμή στη μεταβλητή NMKLTHREADS.

Βήμα 2

Αφού το υβριδικό μοντέλο εκπαιδεύτηκε με ένα κρυφό επίπεδο, αλλάξαμε τη δομή του προσθέτοντας ένα ακόμα επίπεδο με την εντολή HHEd και επαναλαμβάνοντας το Βήμα 1 και 2 μέχρι το μοντέλο να φτάσει το ικανοποιητικό μέγεθος.

HHEd -H dnn/1layer/model -M dnn/2layers/ addlayer2.hed hmm_mono/monolist

4.1.4.2 Στάδιο fine-tuning

Ολοκληρώνοντας το στάδιο του pre-train, προχωρήσαμε σε αυτό του fine-tuning κατά το οποίο βελτιώνονται οι τιμές των βαρών του νευρωνικού δικτύου. Συγκεκριμένα, όταν η ακρίβεια στο σύνολο των δεδομένων, που αναφέρονται στο αρχείο dnn.hv.scp, είναι μεγαλύτερη από 0.001 x 100%, ο ρυθμός εκμάθησης μειώνεται και αν η ακρίβεια μειώνεται εξίσου, το στάδιο αυτό ολοκληρώνεται. Σαν συνάρτηση ενεργοποίησης χρησιμοποιήθηκε η σιγμοειδής. Η εντολή που χρησιμοποιήθηκε είναι ανάλογη με την παράγραφο 4.1.4.1. Η μόνη διαφορά έγκειται στο αρχείο ρυθμίσεων:

HNTrainSGD -C config.basic -C config.finetune -H dnn/4layer/model -M dnn/finetune/ -S dnn.train.scp -N dnn.hv.scp -l LABEL -I mono.aligned.mlf hmm_mono/monolist

όπου, το config.finetune ήταν:

```
MINIBATCHSIZE = 200
MINEPOCHNUM = 10
MAXEPOCHNUM = 15
LEARNRATE = 0.01
CRITERION = XENT
LEARNRATE = 0.001
HIDDENACTIVATION = SIGMOID
OUTPUTACTIVATION = SOFTMAX
NMKLTHREADS = 4
```

Σχημα 4.9: Περιεχόμενο αρχείου config.finetune.

4.1.5 Αποκωδικοποίηση φωνημάτων

Πρωτού συνεχίσουμε, πρέπει να αναφερθεί ότι θα παρουσιαστεί μόνο η διαδικασία που μας οδήγησε στα αποτελέσματα, καθώς αυτά αναλύονται στο Κεφάλαιο 5. Μετά το τέλος της εκπαίδευσης των μοντέλων μας, ορίσθηκε η γραμματική που θα παράγει τις λέξεις-γεγονότα. Το βήμα αυτό είναι υποχρεωτικό καθώς το HTK έχει διαμορφωθεί ώστε να χρησιμοποιείται σε φωνητική αναγνώριση όπου λέξεις είναι το αντικείμενο που πρέπει να εντοπισθεί και να κατηγοριοποιηθεί. Σε αυτό τον κλάδο της αναγνώρισης προτύπων τα μοντέλα Markov αποτελούν την πιο αποδοτική λύση. Στη δική μας περίπτωση, δεν έχουμε λέξεις αλλά γεγονότα. Συνεπώς, η γραμματική που ορίστηκε παρήγαγε τέτοια γεγονότα, είτε μεμονωμένα, είτε ακολουθίες.

Η γραμματική που παράγει μεμονωμένα γεγονότα ορίζεται πολύ εύκολα όπως στο Σχήμα 4.11.

```
S: EVENTS
```

Σχημα 4.10: Ορισμός της γραμματικής στο αρχείο smartplaces.grammar

και το λεξιλόγιο περιγράφει με ποια φωνήματα μπορούν να αντικατασταθούν τα γεγονότα.

```
% EVENTS
KNOCK kn
DOORSLAM ds
STEPS st
CHAIRMOVING cm
SPOON cl
PAPERWORK pw
KEYJINGLE kj
KEYBOARD kt
PHONE pr
APPLAUSE ap
COUGH co
SPEECH spe
SILENCE si
```

Σχημα 4.11: Ορισμός του λεξιλογίου στο αρχείο smartplaces.voca

Η παραγωγή προτάσεων ορίζεται στο αρχείο **gram**:

```
$events= KNOCK | DOORSLAM | STEPS | CHAIRMOVING | SPOON | PAPERWORK |
KEYJINGLE | KEYBOARD | PHONE | APPLAUSE | COUGH | SPEECH | SILENCE;
($events)
```

Σχημα 4.12: Περιεχόμενο αρχείου gram

Με τη δημιουργία των παραπάνω αρχείων, κλήθηκε η εντολή του HTK:

HParse gram wdnet

δημιουργώντας έτσι το αρχείο δικτύου λέξεων **wdnet** που είναι χρήσιμο στη διαδικασία της αποκωδικοποίησης.

Το τελευταίο αρχείο που χρειαζόταν ήταν το λεξικό (**lexicon**). Το λεξικό αποτελείται από όλες τις λέξεις ή συνδυασμό αυτών που υφίστανται σε μία γλώσσα. Το HTK εξ' ορισμού περιέχει το λεξικό του voxforge όπου είναι καταγεγραμμένες η πλειονότητα των αγγλικών λέξεων. Στο δικό μας πρόβλημα όμως, που ασχολείται με ακουστικά γεγονότα, τα φωνητικά δεδομένα δεν υφίστανται. Γίνεται λοιπόν αντιληπτό, πως το αρχείο lexicon θα έχει το ίδιο περιεχόμενο με το smartplaces.voca.

Η εντολή που χρησιμοποιήθηκε για την αποκωδικοποίηση είναι:

HVite -A -D -T 1 -H macros -H epoch11/dnnDef -C config.basic -S decodeData.scp -l '*' -i decoding/recOut.mlf -w wdnet -p 0.0 -s 5.0 voxforge_lexicon tiedlist

το οποίο δίνει σαν έξοδο το αρχείο recOut.mlf. Η χρήση του αναλύεται καλύτερα στο Κεφάλαιο 5. Παράδειγμα του περιεχομένου του παραθέτεται στο Σχήμα 4.13. Τα νούμερα αντιπροσωπεύουν τη χρονική στιγμή εντοπισμού, τη διάρκεια του αναγνωρισμένου γεγονότος και τη λογαριθμική πιθανότητά του.

```
#!MLF!#
"*/S08-eventid-0.rec"
0 2200000 ds -803.827332
.
"*/S08-eventid-1.rec"
0 3600000 ds -1316.290039
.
"*/S08-eventid-2.rec"
0 4400000 ds -1754.076294
.
"*/S08-eventid-3.rec"
0 2200000 ds -809.735901
.
"*/S08-eventid-4.rec"
0 4100000 ds -1460.559570
.
"*/S08-eventid-5.rec"
0 10000000 kn -3700.865234
.
"*/S08-eventid-6.rec"
0 3300000 ds -1223.540161
...and so on...
```

Σχημα 4.13: Περιεχόμενο αρχείου recout.mlf

Ακολουθήθηκαν δύο οδοί κατά την διαδικασία της κατηγοριοποίησης:

- Το σύνολο εκπαίδευσης και αξιολόγησης αποτελούταν από μεμονωμένα γεγονότα-ακουστικά αρχεία (isolated training-testing).

- Το σύνολο εκπαίδευσης αποτελούταν από μεμονωμένα γεγονότα, ενώ σαν δεδομένα δοκιμών χρησιμοποιήθηκε ολόκληρο το αρχείο της συνεδρίας 8 (embedded testing).

Για τη λήψη των αποτελεσμάτων, χωρίς όμως να παρουσιάζονται εδώ, κλήθηκε η εντολή του HTK:

HResults -I testref.mlf tiedlist recout.mlf

Σε αυτό το σημείο τελειώνει η περιγραφή της πορείας που ακολουθήσαμε στο HTK.

4.2 SoX

Το SoX αποτελεί μία πλατφόρμα μέσω γραμμής εντολών που επεξεργάζεται αρχεία ήχου ή τα μετατρέπει σε διαφορετικά φορμα. Επιπλέον, μπορεί να εφαρμόσει διάφορες τροποποιήσεις όπως να πραγματοποιήσει ηχογράφηση, να αλλάξει το ρυθμό δειγματοληψίας ή να τμηματοποιήσει και να συγχωνέψει τα ακουστικά αρχεία. Όσον αφορά τη διπλωματική εργασία, το SoX χρησιμοποιήθηκε στην τμηματοποίηση των ακουστικών αρχείων ώστε κάθε ένα να αποτελεί ένα συμβάν. Επίσης, έγινε αλλαγή του ρυθμού της δειγματοληψίας στα 16kHz. Τέλος από .wv τα αρχεία μετατράπηκαν σε φορμα .wav. Χρειάζεται επίσης να σημεωθεί πως οι εντολές του SoX συσσωρεύτηκαν σε script αρχεία Python με σκοπό την εφαρμογή τους σε ολόκληρη τη βάση δεδομένων.

Οι εντολές που χρησιμοποιήθηκαν ήταν οι εξής:

- ανάκληση πληροφορίας σχετικά με το αρχείο, sox *filename* -n stat.
- αλλαγή του ρυθμού δειγματοληψίας, sox *input.wav* -r 16000 *output.wav* .
- εξαγωγή τμήματος από ακουστικό αρχείο, sox *input output* trim *start duration*.

4.3 Σημειώσεις

Ο κάθε χρήστης δεν χρειάζεται να προβληματιστεί με τον κώδικα υλοποίησης, απλά να χρησιμοποιήσει σωστά τις κλήσεις των συναρτήσεων όπως ορίζει το API. Παρόλο που το HTK αποτελεί ένα καινοτόμο εργαλείο, δεν ανανεώνεται εύκολα με αποτέλεσμα άλλα λογισμικά όπως το Tensorflow της Google να γίνονται πιο ευρέως διαδεδομένα. Τέλος, το HTK μπορεί να μεταφορτωθεί από το http://htk.eng.cam.ac.uk/.

Κεφάλαιο 5

Πειράματα & Αποτελέσματα

5.1 Πειραματικό πλαίσιο

Για την πραγματοποίηση των πειραμάτων σχετικά με το πρόβλημα της κατηγοριοποίησης γεγονότων χρησιμοποίηθηκε το εργαλείο HTK σε συνδυασμό με τη βάση δεδομένων UPC-TALP Multimodal που περιγράψαμε στο Κεφάλαιο 3. Συγκεκριμένα, με το HTK δημιουργήσαμε ένα υβριδικό νευρωνικό δίκτυο από ένα υπάρχων βελτιστοποιημένο HMM μοντέλο, επεκτείναμε και εκπαιδεύσαμε το δίκτυο και έπειτα το αξιολογήσαμε ως προς τη δυνατότητα αναγνώρισης των γεγονότων με τον αποκωδικοποιητή Viterbi. Κατά την διαδικασία της εκπαίδευσης χρησιμοποιήθηκαν μεμονωμένα αρχεία ήχου που το καθένα περιέγραφε ένα γεγονός. Σε αντίθετη περίπτωση, στην αποκωδικοποίηση χρησιμοποιήθηκαν τόσο μεμονωμένα αρχεία ήχου, όσο και ολόκληρο το ακουστικό αρχείο της συνεδρίασης 8. Τα χαρακτηριστικά που χρησιμοποιήθηκαν για να ομαδοποιήσουν τα γεγονότα είναι οι συντελεστές DELTA-DELTA-MFCC. Επίσης, πρέπει να επισημανθεί ότι από τις 8 συνολικά συνεδρίες, οι πρώτες 7 αποτέλεσαν υλικό εκπαίδευσης των ταξινομητών, ενώ η συνεδρία 8 χρησιμοποιήθηκε για την κατηγοριοποίηση των γεγονότων. Τέλος, χρησιμοποιήθηκαν και τα 24 κανάλια.

Κατά τη διάρκεια των πειραμάτων χρησιμοποιήθηκε ένα υπολογιστικό σύστημα τύπου desktop. . Η ολοκληρωμένη διαδικασία της εκπαίδευσης και του ελέγχου αποτελεσμάτων ταξινόμησης σε βαθιά νευρωνικά δίκτυα διήρκησε στο σύνολό της περίπου 60-90 λεπτά ανάλογα με τις παραμέτρους που είχαν δοκιμαστεί. Πρέπει να σημειωθεί ότι χρησιμοποιήθηκαν οι μαθηματικές βιβλιοθήκες MKL της Intel, που επιτρέπουν τον πολυνηματισμό. Τα χαρακτηριστικά του μηχανήματος είναι:

- Ubuntu 16.04 LTS 64bit
- 12GB DDR3 RAM @1666MHz
- Intel Core i7 870@2.93GHz

- GeForce GTX 750Ti
- 60GB Kingston SSD

Τέλος, Ο πίνακας 5.1 παρουσιάζει τις 13 κλάσεις στις οποίες βασίστηκαν τα πειράματα, με 13η να είναι αυτή της ησυχίας.

Ηχητικό Γεγονός	Σύμβολο
Χτύπημα Πόρτας	kn
Κλείσιμο Πόρτας	ds
Βήματα	st
Μετακίνηση Καρέκλας	cm
Κουτάλι/Κουδούνισμα φλιτζανιού	cl
Τύλιγμα Χαρτιού Εργασίας	pw
Ήχος Κλειδιών	kj
Ήχος Πληκτρολογίου	kt
Ήχος Τηλεφώνου/Μουσική	pr
Χειροκρότημα	ap
Βήχας	co
Ομιλία	sp
Ησυχία	si

Πινακας 5.1: Σύνολο κλάσεων της UPC-TALP Multimodal Database

5.2 Μετρικές αξιολόγησης

Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιήθηκαν δύο βασικές μετρικές. Η πρώτη μετρική είναι ανάλογη του WER στον κλάδο της Αυτόματης Αναγνώρισης Ομιλίας, όπου οι λέξεις της γραμματικής αναφέρονται στα ακουστικά γεγονότα. Όμως, αυτή η μετρική δεν λαμβάνει υπόψιν τη χρονική διάρκεια του γεγονότος, αλλά την ακριβή χρονική αντιστοίχιση της πρόβλεψης του ταξινομητή σε σχέση με το ground truth. Συγκεκριμένα, απευθυνόμαστε σε αυτή ως *ποσοστό λάθους αναγνώρισης ακουστικών γεγονότων* acoustic event error rate ή AEER) και χρησιμοποιήθηκε κατά τη διάρκεια της εργασίας όταν τα δεδομένα ήταν μεμονωμένα χωρίς επικαλύψεις. Στη περίπτωση που η ταξινόμηση πραγματοποιήθηκε με αλληλουχία γεγονότων χρησιμοποιήθηκε η μετρική *ποσοστό λάθους ταξινόμησης σε επίπεδο πλαισίου* (frame misclassification rate ή FMR). Η ακρίβεια ενός ταξινομητή [19] σε σχέση με το λόγο σφάλματος λέξης δίνεται από τον ακόλουθο τύπο.

$$Acc = 1 - AEER \tag{5.1}$$

Πιο συγκεκριμένα, το AEER ορίζεται ως ακολούθως

$$AEER = \frac{S + D + I}{N} \tag{5.2}$$

ή

$$AEER = \frac{S + D + I}{S + D + C} \tag{5.3}$$

όπου

- S είναι ο αριθμός των αντικαταστάσεων
- D είναι ο αριθμός των διαγραφών
- I είναι ο αριθμός των εισαγωγών
- C είναι ο αριθμός των σωστών ταξινομημένων παρατηρήσεων
- N είναι το πλήθος των παρατηρήσεων

Τα παραπάνω σύμβολα συνεισφέρουν στον υπολογισμό κάποιων επιμέρους μετρικών λάθους που χρησιμοποιούνται ευρέως στην αναγνώριση ήχων και γεγονότων. Πιο αναλυτικά ορίζουμε ως:

- **Λάθος εισαγωγής (insertion error)**. Το λάθος που υπολογίζεται όταν ο ταξινομητής αναγνωρίζει περισσότερα γεγονότα από όσα έπρεπε.
- **Λάθος διαγραφής (deletion error)**. Αποτελεί ουσιαστικά την αντίθετη περίπτωση από το λάθος εισαγωγής.
- **Λάθος αντικατάστασης (substitution error)**. Αποτελεί την πιο απλή μορφή λάθους καθώς αναφέρεται στο λάθος του ταξινομητή να κατηγοριοποιήσει το γεγονός σωστά. Για παράδειγμα, ενώ έπρεπε να αναγνωριστεί το γεγονός βήματα, τελικά αναγνωρίστηκε χειροκρότημα.

5.3 Αποτελέσματα μετρήσεων HTK

Στο παρών τμήμα του κεφαλαίου θα παρουσιαστούν τα αποτελέσματα των πειραμάτων που πραγματοποιήθηκαν πάνω στη βάση δεδομένων. Αρχικά, γίνεται αναφορά στα αποτελέσματα του εργαλείου HTK τόσο με τη χρήση μεμονομένων γεγονότων, όσο και με ακολουθίες συμβάντων. Η διαδικασία που οδήγησε στη λήψη των αποτελεσμάτων περιγράφεται αναλυτικά στο Κεφάλαιο 4, ενώ μπορούν να βρεθούν και σύνδεσμοι στο τμήμα *Παραρτήματα* σε βασικά τμήματα κώδικα.

Τα χαρακτηριστικά που χρησιμοποιήθηκαν, όπως περιγράφεται και στο Κεφάλαιο 2, ήταν οι συντελεστές DELTA-DELTA-MFCC. Μέσα από αυτά τα αποτελέσματα γίνεται μία εκτίμηση της απόδοσης των επιλεγμένων χαρακτηριστικών αλλά και των εργαλείων όσον αφορά την αναγνώριση ακουστικών γεγονότων. Στην ενότητα αυτή θα γίνει παρουσίαση των εκτιμήσεων του μοντέλου DNN-HMM χρησιμοποιώντας το εργαλείο HTK. Όπως προαναφέρθηκε, τα ηχητικά κομμάτια της συνεδρίας 8 αποτελούν τα στοιχεία αναφοράς του ταξινομητή.

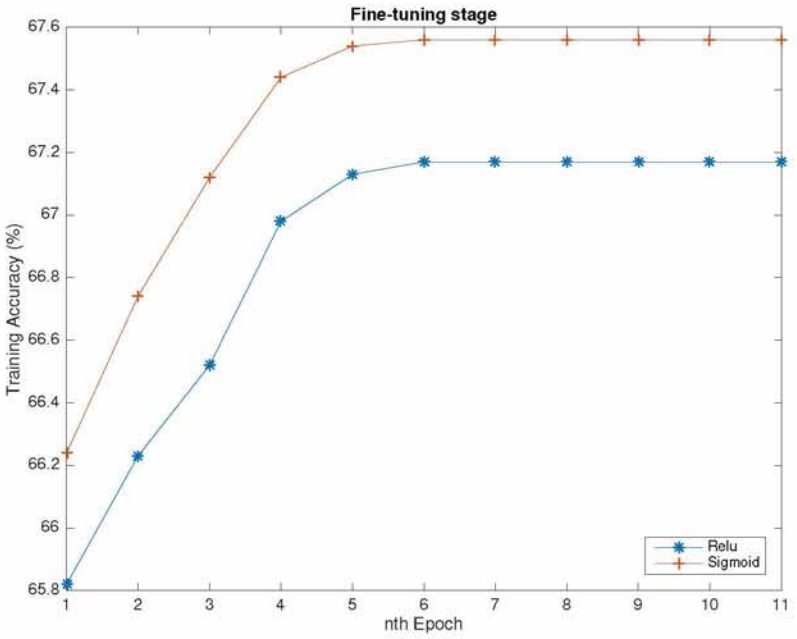
Αρχικά, δοκιμάστηκαν τα μεμονωμένα κομμάτια ήχου (κάθε κομμάτι ήχου αποτελεί ένα συμβάν) και ως validation data η συνεδρία 8 τμηματοποιήθηκε με την ίδια μέθοδο. Δοκιμάστηκαν διάφορες τιμές για την παράμετρο *λάθος εισαγωγής λέξης* (word insertion penalty ή WIP). Καθώς αυξάνεται η μεταβλητή αυτή, η γραμματική που παράγει τις λέξεις-συμβάντα υποχρεώνεται να έχει σαν έξοδο μεμονωμένες λέξεις. Αυτή η παραμετρικοποίηση βοηθάει στην κατακόρυφη μείωση του λαθους εισαγωγής και διαγραφής με αποτέλεσμα την καλύτερη αξιολόγηση του ταξινομητή σε μεμονωμένα γεγονότα ή γνωστό ως isolated validation.
Έχοντας σαν γνώμονα της διαδικασία που pre-train και του fine-tuning, στην έξοδο του δεύτερου η ακρίβεια στην εκπαίδευση ενός μοντέλου με 4 κρυφά επίπεδα, με χρήση SIGMOID, έφτασε το 67.56%, ενώ με RELU στο 67.17%.

```
Epoch 11 *************
Processing training set...
Train accuracy = 67.56% Training time cost = 1786.60s
Processing validation set...
Validation Accuracy = 66.52%
Validation time cost = 75.86s
```

Σχημα 5.1: Αποτέλεσμα fine-tuning με SIGMOID

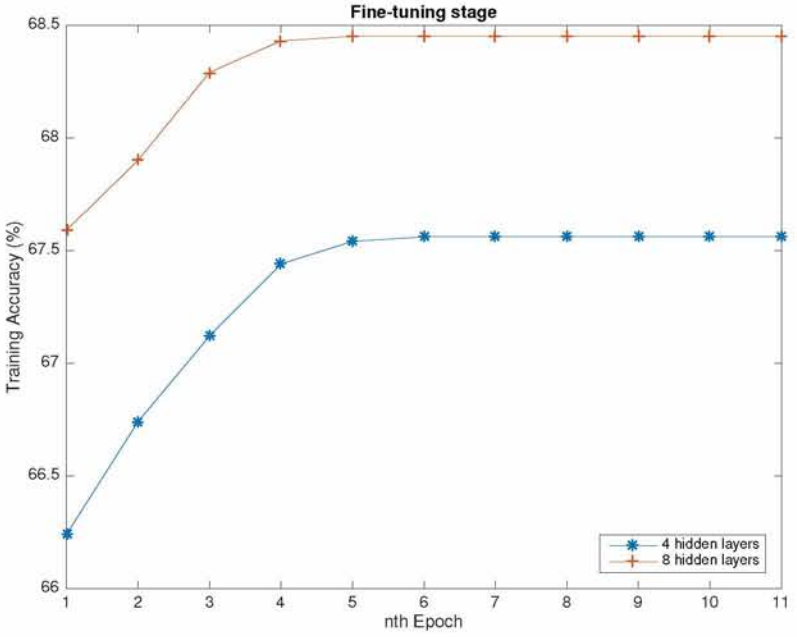
```
Epoch 11 *************
Processing training set...
Train accuracy = 67.17% Training time cost = 1855.44s
Processing validation set...
Validation Accuracy = 66.13%
Validation time cost = 77.06s
```

Σχημα 5.2: Αποτέλεσμα fine-tuning με RELU



Σχημα 5.3: Πρόοδος στην ακρίβεια της εκπαίδευσης με χρήση δύο διαφορετικών συναρτήσεων ενεργοποίησης.

Αφού λοιπόν η sigmoid μας έδωσε καλυτερα αποτελέσματα, επιλέγουμε το ανάλογο μοντέλο. Τέλος επαναλαμβάνουμε την ίδια διαδικασία για ένα νευρωνικό δίκτυο με 8 κρυφά επίπεδα, όπου η ακρίβεια εκπαίδευσης έφτασε το 68.48%. Πρέπει να αναφερθεί ότι δοκιμάστηκαν και νευρωνικά δίκτυα με 16 και 32 κρυφά επίπεδα, αλλά επειδή δεν ήταν αποδοτικότερα, δεν παρουσιάζονται στην εργασία. Η διαφορά των δύο πειραμάτων φαίνεται στο σχήμα 5.4.



Σχημα 5.4: Η επιπλέον προσθήκη 4 επιπέδων βελτίωσε την ακρίβεια της εκπαίδευσης.

Για τη διάτρεξη και ολοκλήρωση του σταδίου της αποκωδικοποίησης χρειαστήκαμε τα παρακάτω αρχεία:

- Το μοντέλο του νευρωνικού δικτύου στο τέλος του 11ου epoch (αρχείο dnnDef).
- Τις παραμέτρους του DNN-HMM (αρχείο config.basic).
- Τα γεγονότα που αναγνώρισε το μοντέλο σε κάθε αρχείο ξεχωριστά (αρχείο recOut.mlf).
- Το FSM και το λεξιλόγιο της γλώσσας μας (αρχείο wdnet και lexicon).
- Τη λίστα με τις καταγεγραμμένες κλάσεις (αρχείο hmmList).

Ο έλεγχος απόδοσης του μοντέλου πραγματοποιήθηκε πάνω στα μεμονωμένα αρχεία του 1ου μικροφώνου της συνεδρίας 8 και στο σύνολο εκπαίδευσης με τη συνάρτηση HVite, όπως παρουσιάζεται στο Κεφάλαιο 4. Επίσης τα μοντέλα που πειραματιστήκαμε ήταν 4 επίπεδα x 128, 4 επίπεδα x 256, 8 επίπεδα x 256. Παρακάτω παρουσιάζονται τα αποτελέσματα ύστερα από την εκτέλεση της εντολής:

HResults -I s08mic1.mlf hmmList recOut.mlf,

όπου, **s08mic1.mlf**, είναι το ground truth, **hmmList** περιέχει τις κλάσεις και το **recOut.mlf** περιέχει όλα τα γεγονότα που αναγνώρισε το νευρωνικό δίκτυο.

HTK		
DNN		GMM-HMM
Layers x Nodes	Correct Words (%)	8 HMM states Correct Words (%)
4x128	78.56	
4x256	78.74	93.73
8x256	78.41	

Σχημα 5.5: Πρόοδος της απόδοσης του ταξινομητή DNN-HMM σε σχέση με τον GMM-HMM σε αποκωδικοποίηση μεμονωμένων γεγονότων.

Όπως παρατηρούμε από το Σχήμα 5.5 η απόδοση του ταξινομητή DNN δεν είναι ανάλογη του προηγούμενου γκαουσιανού ταξινομητή, παρόλο που το υβριδικό νευρωνικό δίκτυο έχει χτιστεί με βάση το καλύτερο GMM-HMM.

Κλείνοντας την μεμονωμένη αποκωδικοποίηση πρέπει να γίνει ξεκάθαρος ο τρόπος με τον οποίο εξήχθησαν τα πραπάνω αποτελέσματα από το HTK. Κατά την εκτέλεση του αποκωδικοποιητή Viterbi, το HTK αποθηκεύει δυναμικά στο αρχείο recOut.mlf το γεγονός που αναγνώρισε ο ταξινομητής σε κάθε ηχητικό απόσπασμα της συνεδρίας 8. Παράδειγμα

αυτού του αρχείου μπορεί να βρεθεί στο τμήμα *Παραρτήματα* στο τέλος της διπλωματικής εργασίας. Έχοντας λοιπόν δημιουργήσει αυτό το αρχείο, το HTK συγκρίνει το ground truth που του έχει ορισθεί στα αρχεία ετικετών με το περιεχόμενο του recOut.mlf. Επειδή όμως τα δεδομένα μας αποτελούνται από μεμονωμένα ακουστικά συμβάντα, το λάθος εισαγωγής και διαγραφής πρέπει να τείνει στο 0, όπως και συμβαίνει, αφού με την διάτρεξη διάφορων πειραμάτων οι τιμές τους παρέμειναν στο 1.

Σε δεύτερη φάση δοκιμάσαμε το σύστημα του HTK σε ενσωματωμένες δοκιμές (embedded testing). Επίσης, για να ενισχύσουμε το μοντέλο της *ησυχίας* (si) εξαγάγαμε περισσότερα κομμάτια ήχου από τα ήδη υπάρχοντα. Αυτό επιτεύχθηκε αφού στις πρώτες 7 συνεδρίες εμφανίζονται χρονικά κενά μεταξύ του τέλους ενός γεγονότος και της αρχής του επόμενου. Υπολογίζοντας έτσι αυτή τη χρονική διαφορά με τη χρήση του εργαλείου SoX, σε κάθε συνεδρία αντιστοιχούσαν τουλάχιστον 150 κομμάτια μικρού μήκους που δεν ανήκαν σε κάποια από τις 12 κλάσεις. Χρειάζεται επίσης να σημειωθεί ότι έχοντας (embedded testing), ως σημείο αναφοράς για την αξιολόγηση του ταξινομητή χρησιμοποιήθηκε ολόκληρο το ακουστικό αρχείο της συνεδρίας 8. Μολαταύτα, αυτό το ακουστικό αρχείο θα μπορούσε να τμηματοποποιηθεί σε 4 επιμέρους κομμάτια ξανά με τη χρήση του εργαλείου SoX. Κάτι τέτοιο όμως δεν πραγματοποιήθηκε αφού το κόψιμο των ακουστικών αρχείων θα επέφερε απώλεια σε ακουστικά παράθυρα. Όλα τα προαναφερθέντα στοιχεία διατυπώθηκαν σαν εισαγωγή για την περιγραφή των αποτελεσμάτων αλλά και της μετρικής FMR που αναφέρθηκε στο υποκεφάλαιο 5.2.

Πριν προχωρήσουμε στην περιγραφή εξαγωγής των αποτελεσμάτων πρέπει να αναφερθεί ότι το αρχείο recOut.mlf μετά τη διαδικασία του decoding είχε σχεδόν την ίδια μορφή με αυτή των μεμονωμένων πειραμάτων, περιέχοντας όμως ένα αρχείο, της συνεδρίας 8, διατρέχοντάς το χρονικά και αναγνωρίζοντας τα γεγονότα σε επίπεδο ms. Με σκοπό λοιπόν τη χρήση αυτής της μετρικής, γράφτηκε κώδικας σε python που πραγματοποιούσε τις εξείς ενέργειες:

1. Ανοίγει το αρχείο S08.csv και αναλύει τα γεγονότα σε πλαίσια. Για παράδειγμα, αν ένα γεγονός διαρκεί από 0 έως 2,7 δευτερόλεπτα, τότε υπάρχουν 270 παράθυρα των 10ms που περιγράφουν αυτό το γεγονός. Στο σύνολο της διάρκειας του ακουστικού αρχείου της συνεδρίας 8, υπολογίζονται 149991 παράθυρα. Συνεπώς, από το αρχείο εξόδου του HTK εξαγάγαμε το ανάλογο πλήθος παραθύρων. Τα χρονικά κενά προφανώς συμπληρώθηκαν με το γεγονός της σιωπής (si).

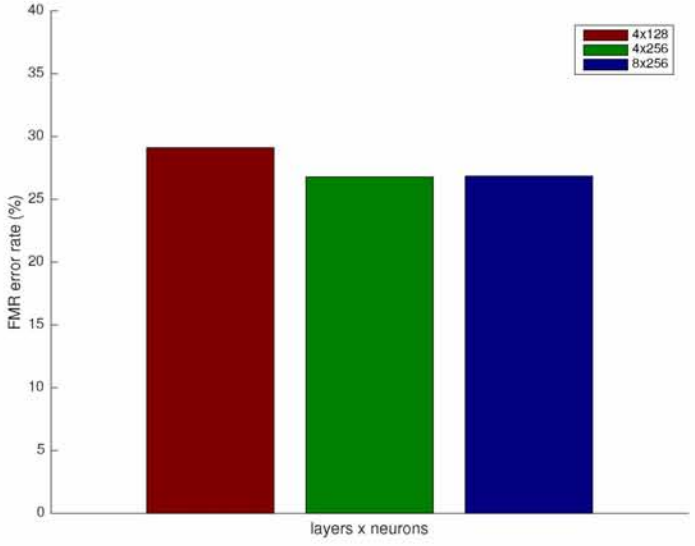
2. Ανοίγει το αρχείο εξόδου recOut.mlf και επαναλαμβάνει την ίδια διαδικασία.

3. Συγκρίνει τα παράθυρα 1-1 και υπολογίζει τον λόγο $\frac{\#matched\_frames}{\#total\_frames}$. Συμπερασματικά, έχει καταστεί σαφές ότι αυτός ο λόγος περιγράφει το ποσοστό λάθους ταξινόμισης σε επίπεδο πλαισίου.

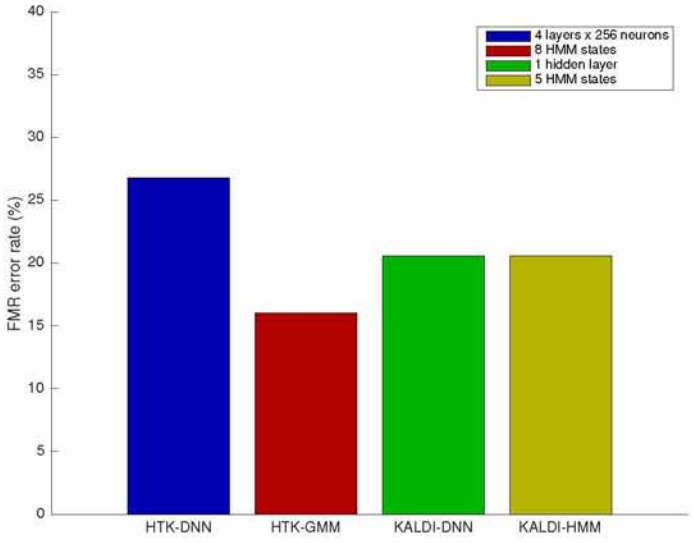
Έχοντας συνθέσει αυτόν τον μικρό αλγόριθμο καλούμε τις εντολές του HTK που αναφέρθηκαν στο Κεφάλαιο. Στο πλαίσιο αυτό κατανοούμε ότι δεν μπορούμε να χρησιμοποιήσουμε μία γραμματική που παράγει μία λέξη, αφού το ακουστικό αρχείο για την αξιολόγηση του ταξινομητή αποτελείται από μία πρόταση με την ακολουθία όλων των γεγονότων που συνέβησαν στην συνεδρία 8. Συνεπώς, μέσα από τις δοκιμές διαφόρων τιμών, όπως παρουσιάζεται στο Σχήμα 5.6, καταλήγουμε στο βέλτιστο αποτέλεσμα που μπορεί να δώσει το HTK. Για λόγους σύγκρισης, αναφέρουμε τα αποτελέσματα των γκαουσιανών μοντέλων μίξης σε HTK και των νευρωνικών δικτύων του εργαλείου Kaldi από το προηγούμενο σχέδιο εργασίας "Αναγνώριση ακουστικών γεγονότων με βαθιά νευρωνικά δίκτυα" [8].

HTK			KALDI	
DNN		GMM-HMM	DNN	GMM-HMM
Layers x Nodes	FER(%)	8 HMM states FER(%)	1 hidden layer FER (%)	5 HMM States FER (%)
4x128	29.09			
4x256	26.77	16.02	20.58	20.58
8x256	26.85			

Σχημα 5.6: Απόδοση ταξινομητών σε επίπεδο ενσωματωμένων γεγονότων.



Σχημα 5.7: Γραφική αναπαράσταση αποτελεσμάτων του Σχήματος 5.6



Σχημα 5.8: Συγκεντρωτικός πίνακας αποτελεσμάτων

Με αναφορά τον παραπάνω πίνακα, παρατηρείται πως το ποσοστό λάθους για το υβριδικό μας μοντέλο αγγίζει το 26.77%, που είναι αρκετά ικανοποιητικό αν αναλογιστούμε ότι καθώς τα γεγονότα συμβαίνουν στον χρόνο, υπάρχουν παύσεις που μπορεί να λειτουργούν ως εμπόδια στην ακριβή οριοθέτηση του γεγονότος μέσα σε αυτόν. Το αποτέλεσμα αυτό έγινε εφικτό με την καλή εκπαίδευση του μοντέλου σιωπής. Γίνεται, επομένως, εύκολα αντιληπτό ότι η ενέργειά μας πάνω στα δεδομένα εκπαίδευσης βελτίωσε σε μεγάλο βαθμό την απόδοση του DNN-HMM ταξινομητή.

Κεφάλαιο 6

Συμπεράσματα

6.1 Συνεισφορά της διπλωματικής εργασίας

Μέσα από την εκπόνηση αυτής της μεταπτυχιακής εργασίας πραγματοποιήθηκε συστηματική μελέτη της απόδοσης των βαθιών νευρωνικών δικτύων στο πρόβλημα της αναγνώρισης και κατηγοριοποίησης ακουστικών γεγονότων που λαμβάνουν χώρα σε ένα δωμάτιο συνεδρίασης. Επίσης, έγινε σύγκριση μεταξύ της απόδοσης του DNN ταξινομητή και των κρυφών μοντέλων Markov. Μέσα από τα πειραματικά αποτελέσματα που παρουσιάστηκαν στο προηγούμενο κεφάλαιο, η επιστημονική συμβολή της διπλωματικής καλύπτει τους εξής άξονες:

- Τη μελέτη της πολυκαναλικής βάσης δεδομένων UPC-TALP που δημιουργήθηκε από την ηχογράφηση 12 διαφορετικών γεγονότων σε ένα έξυπνο περιβάλλον. Επίσης, έγινε επεξεργασία των ακουστικών δεδομένων με τη χρήση του λογισμικού SoX.

- Την εξαγωγή των κατάλληλων χαρακτηριστικών που θα μπορούσαν να περιγράψουν τα δεδομένα μας καλύτερα. Στην αρχή των πειραμάτων, όπως αναφέρουμε και στα προηγούμενα κεφάλαια, χρησιμοποιήθηκαν οι 39 (MFCC, D-MFCC, DD-MFCC).

- Τη χρήση διαφορετικών μετρικών εκτός των προκαθορισμένων, καθώς με τη μελέτη ακουστικών γεγονότων και όχι λέξεων, το word error rate δεν είχε υπόσταση. Έτσι, έγινε χρήση της μετρικής *ποσοστό λάθους ταξινόμησης ακουστικών γεγονότων* (για μεμονομένα συμβάντα) και *ποσοστό λάθους ακολουθίας γεγονότων σε επίπεδο πλαισίου*. Με τον ορισμό αυτών των μετρικών, μπορέσαμε να αξιολογήσουμε καλύτερα τους ταξινομητές σε διαφορετικές φάσεις του προβλήματος. Η μετρική FMR είναι αυτή που παρουσίασε ιδιαίτερο ενδιαφέρον αφού εξέταζε τη συνεδρία 8 σε λεπτομερές επίπεδο χρόνου. Παρόλο που και τα δύο εργαλεία (HTK, Kaldi) χρησιμοποιούν τους ίδιους αλγορίθμους εκπαίδευσης και αποκωδικοποίησης, αυτή η διαφοροποίηση στο αποτέλεσμα πιθανόν οφείλεται στο διαφορετικό σχεδιασμό των εργαλείων παρά στην υλοποίηση των αλγορίθμων.

- Τέλος, η πιο σημαντική συνεισφορά της διπλωματικής αυτής ήταν τα πειράματα με διαφορετικές ρυθμίσεις για τα νευρωνικά δίκτυα, αλλά και η δημιουργία υβριδικών αρχιτεκτονικών ώστε να μελετηθεί καλύτερα η συμπεριφορά τους. Όπως αναφέρουν και στην εργασίας τους οι Hugo Larochelle, Yoshua Bengio κ.α [17], η στρατηγική εκμάθησης του αλγορίθμου οπισθοδιάδοσης δίνει μη-ικανοποιητικές λύσεις για νευρωνικά δίκτυα με πολλά κρυφά επίπεδα.

Σχολιάζοντας τα αποτελέσματα αυτά, φαίνεται ότι τα μοντέλα Markov συμπεριφέρονται καλύτερα στο πρόβλημά μας, απ΄ ότι τα βαθιά νευρωνικά δίκτυα.

6.2 Μελλοντικές ερευνητικές κατευθύνσεις

Μετά τη διεκπεραίωση της παρούσας εργασίας μπορούν να ακολουθηθούν επιπρόσθετοι βελτιωτικοί οδοί σε μελλοντικές έρευνες. Μερικές από αυτές παρουσιάζονται παρακάτω:

- Με σκοπό τη βελτίωση των ακουστικών χαρακτηριστικών θα μπορούσε να γίνει περαιτέρω έρευνα πάνω στη χρήση των MFCC σε συνδυασμό με τις τιμές TDOA. Με αυτή τη διαδικασία θα δημιουργηθούν μεγαλύτερα διανύσματα χαρακτηριστικών που πιθανώς να περιγράφουν καλύτερα τα γεγονότα. Επίσης, σύμφωνα με την βιβλιογραφία η απόδοση του ταξινομητή μπορεί να αυξηθεί με την εφαρμογή μη-φασματικών χαρακτηριστικών με θετικό συνελικτικό πίνακα παραγοντοποίησης μαζί με τα χαρακτηριστικα MFCC[20].

- Ένα άλλο πεδίο μελλοντικής έρευνας αποτελεί ο επαναπροσδιορισμός των αρχείων ετικετών της βάσης δεδομένων UPC-TALP, με σκοπό την επαναξέταση της απόδοσης του συστήματος βαθιών νευρωνικών δικτύων σε αυτή. Συγκεκριμένα, τα γεγονότα που συμβαίνουν στη συνεδρία 8 είναι περισσότερα από τα καταγεγραμμένα και λειτουργούν σαν θόρυβος παρασκηνίου.

- Ακόμη, ένα πεδίο που πρέπει να επικεντρωθούμε είναι ο συνδυασμός των καναλιών, είτε στο στάδιο της εκπαίδευσης του ταξινομητή, είτε στη φάση της κατηγοριοποίησης. Δηλαδή, κάποια γεγονότα μπορεί να έχουν ηχογραφηθεί καλύτερα από κάποιο άλλο μικρόφωνο. Αυτό θα ενίσχυε αρκετά το κάθε μοντέλο-κλάση κατά τη διαδικασία της εκμάθησης του ταξινομητή. Με αυτόν τον τρόπο θα αποφεύγαμε την υπερ-εκπαίδευση, καθώς δεν θα χρησιμοποιούνταν ολόκληρες οι ηχογραφήσεις και απο τα 24 μικρόφωνα.

- Τέλος, κάθώς η βάση δεδομένων εμπεριέχει και οπτικό υλικό θα μπορούσε να πραγματοποιηθεί συνδυασμός της οπτικοακουστικής πληροφορίας. Δηλαδή να επιτευχθεί συνδυασμός τροπικοτήτων (modality fusion) με σκοπό την αύξηση της ακρίβειας των βαθιών νευρωνικών δικτύων σαν ταξινομητές. Παραδείγματα τέτοιων πειραμάτων εμφανίζονται στα πειράματα των Samira Ebrahimi Kahou1, Christopher Pal κ.α

όπου συνδύασαν βαθιά συνελικτικά νευρωνικά δίκτυα (deep convolutional neural networks), DNN και SVM με σκοπό την αναγνώριση κίνησης σε βίντεο [21].

Bibliography

[1] Taras Butko, Fran González Pla, Carlos Segura, Climent Nadeu, and Javier Hernando. Two-source acoustic event detection and localization: Online implementation in a smart-room. In *Signal Processing Conference, 2011 19th European*, pages 1317–1321. IEEE, 2011.

[2] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2012.

[3] Beth Logan et al. Mel Frequency Cepstral Coefficients for Music Modeling. In *ISMIR*, 2000.

[4] Hidden Markov Model figure. Factor graphs: HMM. URL http://www.igi.tugraz.at/lehre/MLA/WS13/MLA_Exercises_2013/node6.html. [Online; accessed 29-August-2015].

[5] Taras Butko, Climent Nadeu Camprubí, et al. Detection of overlapped acoustic events using fusion of audio and video modalities. 2010.

[6] Noam Shabtai. Acoustic scene analysis. 2010.

[7] Andrey Temko, Robert Malkin, Christian Zieger, Dusan Macho, Climent Nadeu, and Maurizio Omologo. Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems. *Cough*, 65(48):5, 2006.

[8] Gerasimos Potamianos Konstantinos Themelis. Acoustic event recognition using deep neural networks. 2015.

[9] Panagiotis Giannoulis, Gerasimos Potamianos, Athanasios Katsamanis, and Petros Maragos. Multi-microphone fusion for detection of speech and acoustic events in smart spaces. In *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pages 2375–2379. IEEE, 2014.

[10] David H Hubel MD John Franklin et al. *Brain and Visual Perception: The Story of a 25-Year Collaboration*. Oxford University Press, 2004.

[11] Steven B Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, 1980.

[12] Bertrand Denis, Jean Côté, and René Laprise. Spectral decomposition of two-dimensional atmospheric fields on limited-area domains using the discrete cosine transform (dct). *Monthly Weather Review*, 130(7):1812–1829, 2002.

[13] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. *The HTK book*, volume 2. Entropic Cambridge Research Laboratory Cambridge, 1997.

[14] Ossama Abdel-Hamid, Li Deng, Dong Yu, and Hui Jiang. Deep segmental neural networks for speech recognition. In *Interspeech*, volume 36, page 70, 2013.

[15] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6): 82–97, 2012.

[16] Sergios Theodoridis, Aggelos Pikrakis, Konstantinos Koutroumbas, and Dionisis Cavouras. *Introduction to Pattern Recognition: A Matlab Approach*. Academic Press, 2010.

[17] Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin. Exploring strategies for training deep neural networks. *The Journal of Machine Learning Research*, 10:1–40, 2009.

[18] A Temko, D Macho, C Nadeu, and C Segura. UPC-TALP database of isolated acoustic events. *Internal UPC report*, 85, 2005.

[19] Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C Gates. An empirical analysis of word error rate and keyword error rate. In *INTERSPEECH*, pages 2070–2073, 2008.

[20] Courtenay V Cotton and Daniel PW Ellis. Spectral vs. spectro-temporal features for acoustic event detection. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pages 69–72. IEEE, 2011.

[21] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM, 2013.